

# **SANDIA REPORT**

SAND2007-0939

Unlimited Release

Printed May 2007

## **Experimental Uncertainty Estimation and Statistics for Data Having Interval Uncertainty**

Scott Ferson, Vladik Kreinovich, Janos Hajagos, William Oberkampf and Lev Ginzburg

Prepared by  
Sandia National Laboratories  
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia is a multiprogram laboratory operated by Sandia Corporation,  
a Lockheed Martin Company, for the United States Department of Energy's  
National Nuclear Security Administration under Contract DE-AC04-94AL85000.

Approved for public release; further dissemination unlimited.

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

**NOTICE:** This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, make any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represent that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof, or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors.

Printed in the United States of America. This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from  
U.S. Department of Energy  
Office of Scientific and Technical Information  
P.O. Box 62  
Oak Ridge, TN 37831

Telephone: (865) 576-8401  
Facsimile: (865) 576-5728  
E-Mail: [reports@adonis.osti.gov](mailto:reports@adonis.osti.gov)  
Online ordering: <http://www.osti.gov/bridge>

Available to the public from  
U.S. Department of Commerce  
National Technical Information Service  
5285 Port Royal Rd.  
Springfield, VA 22161

Telephone: (800) 553-6847  
Facsimile: (703) 605-6900  
E-Mail: [orders@ntis.fedworld.gov](mailto:orders@ntis.fedworld.gov)  
Online order: <http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online>



# **Experimental Uncertainty Estimation and Statistics for Data Having Interval Uncertainty**

Scott Ferson, Vladik Kreinovich, Janos Hajagos, William Oberkampf and Lev Ginzburg

Applied Biomathematics  
100 North Country Road  
Setauket, New York 11733

## **Abstract**

This report addresses the characterization of measurements that include epistemic uncertainties in the form of intervals. It reviews the application of basic descriptive statistics to data sets which contain intervals rather than exclusively point estimates. It describes algorithms to compute various means, the median and other percentiles, variance, interquartile range, moments, confidence limits, and other important statistics and summarizes the computability of these statistics as a function of sample size and characteristics of the intervals in the data (degree of overlap, size and regularity of widths, etc.). It also reviews the prospects for analyzing such data sets with the methods of inferential statistics such as outlier detection and regressions. The report explores the tradeoff between measurement precision and sample size in statistical results that are sensitive to both. It also argues that an approach based on interval statistics could be a reasonable alternative to current standard methods for evaluating, expressing and propagating measurement uncertainties.

## Acknowledgments

We thank Roger Nelsen of Lewis and Clark University, Bill Huber of Quantitative Decisions, Gang Xiang, Jan Beck, Scott Starks and Luc Longpré of University of Texas at El Paso, Troy Tucker and David Myers of Applied Biomathematics, Chuck Haas of Drexel University, Arnold Neumaier of Universität Wien, and Jonathan Lucero and Cliff Joslyn of Los Alamos National Laboratory for their collaboration. We also thank Floyd Spencer and Steve Crowder for reviewing the report and giving advice that substantially improved it. Jason C. Cole of Consulting Measurement Group also offered helpful comments. Finally, we thank Tony O'Hagan of University of Sheffield for challenging us to say where intervals come from. The work described in this report was performed for Sandia National Laboratories under Contract Number 19094. William Oberkampf managed the project for Sandia. As always, the opinions expressed in this report and any mistakes it may harbor are solely those of the authors.

This report will live electronically at <http://www.ramas.com/intstats.pdf>. Please send any corrections and suggestions for improvements to [scott@ramas.com](mailto:scott@ramas.com).

# Contents

Acknowledgments.....	4
Contents .....	5
Figures.....	7
Symbols.....	8
Executive summary.....	9
1 Introduction.....	11
2 Where do interval data come from?.....	15
3 Interval analysis .....	19
3.1 Interval arithmetic.....	19
3.2 Outward-directed rounding.....	20
3.3 Repeated parameter problem .....	20
3.4 Two interpretations of an interval.....	21
3.5 Kinds of interval data sets.....	23
4 Descriptive statistics for interval data.....	28
4.1 Distribution .....	30
4.1.2 Confidence limits on the distribution.....	34
4.1.3 Weighted data .....	35
4.2 Arithmetic mean and similar measures of central tendency .....	36
4.2.1 Arithmetic mean.....	36
4.2.2 Geometric mean.....	37
4.2.3 Harmonic mean.....	37
4.2.4 Weighted mean .....	38
4.3 Median and percentiles .....	38
4.4 No useful estimate for mode.....	39
4.5 Variance and other measures of dispersion .....	41
4.5.1 Variance .....	41
4.5.2 Standard deviation and standard error .....	58
4.5.3 Range and interquartile range .....	59
4.6 Moments .....	60
4.6.1 Moments about the origin.....	61
4.6.2 Central moments of even order.....	61
4.6.3 Central moments of odd order .....	64
4.7 Skewness.....	64
4.8 Confidence intervals .....	66
4.8.1 Inner bounds on confidence limits.....	68
4.8.2 Computing the outer bounds on confidence limits .....	71
4.8.3 Confidence intervals without a normality assumption.....	74
4.8.4 Confidence limits for other statistics .....	75
4.9 Fitting named distributions to interval data .....	75
4.9.1 Method of matching moments .....	76
4.9.2 Fitting distributions by regression .....	81
4.9.3 Maximum likelihood methods.....	83
4.9.4 Other methods for fitting distributions .....	87

5	Inferential statistics with interval data .....	88
5.1	Equality of variances for imprecise data.....	89
5.2	Detecting outliers.....	90
5.3	Linear regression.....	94
6	Computability of interval statistics .....	101
7	Tradeoff between sample size and precision .....	104
8	Methods for handling measurement uncertainty.....	110
8.1	The standard approach .....	110
8.2	Using interval statistics within the standard approach.....	117
8.2.1	Dependencies among sources of uncertainty .....	121
8.2.2	Accounting for incertitude in measurement uncertainty.....	123
8.3	Harmonization.....	125
8.4	Is measurement uncertainty underestimated? .....	127
9	Conclusions and next steps .....	129
9.1	What should we call this? .....	131
9.2	Next steps.....	131
10	Appendix.....	133
11	Glossary .....	134
12	References.....	140

## Figures

Figure 1. Two interpretations about what an interval represents.....	22
Figure 2. Examples of the various kinds of interval data sets .....	26
Figure 3. Relationships among the different kinds of interval data sets.....	27
Figure 4. Hypothetical interval data set ‘skinny’ .....	29
Figure 5. Hypothetical interval data set ‘puffy’ .....	29
Figure 6. Empirical distributions for the skinny and puffy data sets .....	32
Figure 7. Empirical distributions under the equidistribution hypothesis.....	33
Figure 8. Confidence limits on the distribution functions .....	35
Figure 9. The medians for interval data sets .....	39
Figure 10. The most-overlap statistic.....	40
Figure 11. Scalar values within the skinny data set leading to extremal variances .....	42
Figure 12. Zones for computing the lower bound on variance for the puffy data .....	45
Figure 13. Variance is a quadratic function of each data value .....	48
Figure 14. Scalar values within the puffy data extremizing variance.....	50
Figure 15. Dispersive distributions leading to conservative upper bounds on variance..	58
Figure 16. Interquartile ranges for interval data sets .....	60
Figure 17. Confidence limits on the means .....	72
Figure 18. Distribution-free confidence intervals on the mean .....	75
Figure 19. Exponential distributions fitted to the skinny and puffy data.....	76
Figure 20. Normal distribution and an enclosing p-box .....	80
Figure 21. Normal distributions fitted by the method of matching moments.....	81
Figure 22. Regressions of $X$ against normal scores .....	83
Figure 23. Fitting an exponential distribution by maximum likelihood .....	84
Figure 24. Uncertainty growing until narrowed intervals intersect .....	94
Figure 25. Envelope of regression lines for interval data .....	96
Figure 26. An example of an all-constraints regression .....	96
Figure 27. Two reasons a regression may lack a strong signal.....	97
Figure 28. Bivariate linear regressions with uncertainty in one only variable .....	98
Figure 29. Data for the numerical example .....	99
Figure 30. Bivariate interval data in nonparametric linear regression.....	100
Figure 31. Tradeoff between measurement precision and sampling effort .....	105
Figure 32. Isoclines for cost and uncertainty for different experimental designs.....	106
Figure 33. Optimum designs for each of two cost isoclines .....	107
Figure 34. Line of best experimental designs from cheap to expensive .....	107
Figure 35. Standard ISO/NIST models of measurement uncertainty .....	113
Figure 36. Poor approximation for a nonlinear function and large uncertainties .....	114
Figure 37. Measurement uncertainty model compared to the data’s distribution.....	116
Figure 38. Repeated measurements of the speed of light by Michelson et al.....	116
Figure 39. Standard and interval-statistics models for uncertainty.....	117
Figure 40. Propagation of uncertainty about the sum .....	120
Figure 41. Uncertainty about the sum with and without a dependence assumption.....	122
Figure 42. Possible models for measurement uncertainty about a single measurement	125
Figure 43. Measurement uncertainty for the speed of light at different years .....	128

## Symbols

$X$	a scalar (real) value, or realization of a scalar value from within an interval
$x$	an interval
$x_i$	a collection of intervals indexed by an integer $i$
$\underline{x}$	left endpoint of an interval
$\bar{x}$	right endpoint of an interval
$[\underline{x}, \bar{x}]$ , $[A, B]$	intervals specifying their endpoints explicitly
$]A, B[$ , $]A, B]$	an open or partially open interval
$[\underline{B}(X)$ , $\bar{B}(X)]$	a p-box (cf. Ferson et al. 2003)
$(a, b, c, \dots)$	a vector or $n$ -tuple
$\{a, b, c, \dots\}$	a set
$\pm$	plus or minus
$\sim$	is distributed as
$\approx$	is approximately
$\equiv$	is defined to be
$\rightarrow$	implies
$\in$	is an element of
$\notin$	is not an element of
$\subseteq$	is a subset of
$\not\subseteq$	is not a subset of
$\Sigma$	summation
$\cup$	union
$\emptyset$	the empty set
$\infty$	infinity, or simply a very large number
$\Phi$	standard normal cumulative distribution function
$E$	the true scalar value of the arithmetic mean of measurands in a data set
$N$	sample size
$N(m, s)$	normal distribution with mean $m$ and standard deviation $s$
$\mathcal{O}(\cdot)$	on the order of
$P(\cdot)$	probability of
$\mathbb{R}$	real line
$\mathbb{R}^*$	extended real line, i.e., $\mathbb{R} \cup \{-\infty, \infty\}$
$S_M(X)$	the fraction of $N$ values in a data set that are at or below the magnitude $X$

*Note on typography.* Upper case letters are used to denote scalars and lower case letter are used for intervals. (The only major exception is that we still use lower case  $i, j$ , and  $k$  for integer indices.) For instance, the symbol  $E$  denotes the scalar value of the arithmetic mean of the measurands in a data set, even if we don't know its magnitude precisely. We might use the symbol  $e$  to denote an interval estimate of  $E$ . We use underlining and overbars to refer to the endpoints of intervals. For instance,  $\underline{x}$  and  $\bar{x}$  denote the left and right endpoints of an interval  $x$ , so  $\underline{x}$  and  $\bar{x}$  are therefore scalars. We sometimes use an uppercase letter with underlining or an overbar to denote a (scalar) bound on the quantity.



## Executive summary

Verification and validation of simulation models are increasingly important topics in engineering, especially in contexts where model results are used to make significant certifications or inferences but relevant empirical information is sparse or of poor quality. To support realistic and comprehensive validation efforts, improved characterization of measurement uncertainties is needed. This report summarizes recent work to develop algorithms for statistical analysis of data sets containing measurements (or calculation results) that are better represented by intervals than by point estimates. It also suggests a generalization of standard approaches for characterizing experimental uncertainty that carefully distinguishes and keeps separate different forms of uncertainty.

Epistemic uncertainty that is naturally modeled by intervals is called *incertitude*. This kind of uncertainty is characterized by the absence of evidence about the probabilities that the underlying unknown quantity is in any particular subset of the interval. Incertitude arises in several natural empirical settings, including from binning real-valued data into coarse categories, statistical censoring such as non-detects and “never-failed” observations, intermittent observations, and digital instrument readouts. The report describes data sources where intervals rather than point estimates are the natural product and reviews methods for summarizing interval data. It introduces straightforward extensions of elementary statistical methods to explicitly account for incertitude in empirical data. It also classifies different kinds of data sets in terms of the breadth and nature of the intervals and explains how these differences can simplify or complicate the calculations.

Most widely used statistical methods are not designed to handle incertitude, so it is commonly ignored in formal analyses. Yet it would seem that this incertitude should not be neglected, at least when it is large relative to the magnitude of the measurement or relative to conclusions that might depend on the data. Including such uncertainty in analyses requires the development of statistical methods that do not model incertitude as though it were randomness. The report reviews a variety of methods for computing descriptive statistics of collections of point and interval estimates, such as the mean and other measures of central tendency (geometric mean, median, etc.), the standard deviation and other measures of dispersion (variance, interquartile range), confidence limits, moments, covariance and correlation. It also introduces some methods for statistical inference such as testing equality of variances and detecting outliers. The computability of these new methods is an issue because, in general, evaluating statistical summaries of collections of data including intervals can become impractical for large samples sizes. Luckily, however, in several special cases the problems become computationally tractable.

An important practical application for interval statistics is in expressing and propagating measurement uncertainty. Engineers and physicists have long agreed that any empirical measurement must be accompanied by a statement about the measurement’s uncertainty. Traditionally, however, these statements have taken many forms, including significant

digits, plus-or-minus ranges, confidence limits, and pairs of means and dispersion estimates. In the past, there has been a great deal of variation among the forms of these statements, and, as a consequence, broad confusion about how the statement should be interpreted. It is, still today, not always clear whether, for instance, the second term in a measurement reported as “ $12.43 \pm 0.09$ ” refers to the standard deviation, the standard error, the half-width of a range asserted to enclose the value, or something else. The attempt to regularize practice in this area has led, it might be argued, to a premature ossification of the ideas about measurement uncertainty. The standard for characterizing measurement uncertainty currently recommended by the International Organization for Standardization (ISO) and the National Institute for Standards and Technology (NIST) is based on methods that make many assumptions about the measurement process. In particular, these methods are based on a normal (Gaussian) model of uncertainty propagation, which assumes that uncertainties are linearizably small and have symmetric distributions without complex dependencies. Although the standard methods are recommended for use with all measurements, they do not recognize incertitude as distinct from measurement uncertainty that can be well modeled as randomness by probability distributions. This means the standard methods for modeling measurement uncertainty may not be appropriate in empirical situations where incertitude is important. The methods of interval statistics described in this report are offered as an alternative approach that may be useful in such circumstances. This alternative approach is more expressive, distinguishes epistemic uncertainty (lack of knowledge) and aleatory uncertainty (variability), and relaxes restrictive assumptions in the standard methods.

The text includes detailed algorithms for computing a wide variety of descriptive statistics for data with incertitude, but, as a convenience for the reader, technical material that can be skipped without loss for one interested in the main ideas is clearly marked.

# 1 Introduction

The fundamental precept explored in this report is that, in some important empirical situations, data can come to us that have uncertainties that are naturally represented by mathematical intervals and that cannot properly be characterized by any particular probability distribution. These intervals are used to represent measurement uncertainty, or other uncertainties subsequently attached to a measurement during recording, transcription, transmission, communication or manipulation. The wider an interval is, the greater the uncertainty associated with that measurement.

Although an individual measurement in the form of an interval cannot be associated with a probability distribution, a collection of such data together have a distribution, and it seems reasonable to try to apply standard statistical methods to analyze these data even though they lack the apparent precision of point estimates. Several recent papers have explored this idea and have started developing formulas and algorithms necessary to compute basic statistics for interval data\* (e.g., Ferson et al. 2002a,b; 2005; Wu et al. 2003; Kreinovich, et al. 2005b,c; Gioia and Lauro 2005; Xiang et al. 2006). In fact, there has been quite a lot of work on applying statistical methods to data sets containing interval uncertainty, although a great deal of it undoubtedly has not been heralded in papers in the open literature. This is because many analysts would regard it as a routine application of a bounding argument in a statistical context. The exceptions in the last five years have included wide-ranging developments and applications, including pattern recognition and discriminant function analysis (Nivlet et al. 2001a,b), Bayesian methods (Fernández et al. 2001; 2004), regression analysis (Billard and Diday 2000; Manski and Tamer 2002; Marino and Palumbo 2002; 2003), time series analysis (Möller and Reuter 2006), principal components analysis (Lauro and Palumbo 2000; 2003; 2005; Gioia and Lauro 2006; Lauro et al. 2007), factorial analysis (Palumbo and Irpino 2005), outlier detection (Kreinovich et al. 2005a; Neumann et al. 2006), hypothesis testing (Kutterer 2004), classification (Zaffalon 2002; 2005), and cluster analysis (Chavent et al. 2005). Manski (2003) authored the first monograph on the subject of identifying distributions and parameters of these distributions from data sets with interval uncertainty. Although he did not concern himself with sampling theory or the issues of statistical inference, his book provides an excellent and rigorous grounding in the motivations for this work and many of the basic issues.

One of the fundamental ideas of using intervals to model measurement uncertainty is that data differ in *quality*. Some data are pretty good, and some aren't as good. The latter are represented by wider intervals than the former. Some data are very poor because their intervals are very wide, or even vacuous if the data are missing entirely. Faced with reason to believe that data have differing quality, statisticians using conventional methods

---

\*The phrase 'interval data' as used in this report refers to a data set that consists partly or entirely of intervals, rather than exclusively of point values. The phrase should not be confused with another common usage that refers to the metric properties of the scale on which data are measured. In that usage, interval data, as distinguished from nominal, ordinal, and ratio data, are characterized on a scale on which relative difference is meaningful but for which the zero point is arbitrary.

might assign weights to the data values to represent these differences. This approach may often be reasonable, but it produces results that do not outwardly differ in their appearance in any way from those based on data of equal quality. Likewise, in the conventional approach a data set composed of poor estimates might yield an empirical distribution and summary statistics that are indistinguishable in form or apparent reliability from those of another data set of the same size that is composed of excellent, careful measurements. In contrast, interval statistics leads to qualitatively different results that immediately reveal the essential difference in quality of the data sets.

In the extreme case of missing data, conventional analyses typically make use of methods such as replacing values with the average of the available data, or the last value sampled (“hot deck imputation”) to try to minimize bias in the data set (Little and Rubin 1987). These methods are typically only appropriate if the absent data are missing at random, that is, in a way that is unrelated to their magnitudes. Sometimes this may be a reasonable assumption, and sometimes it may not be. Many analysts feel great discomfort with this assumption, even though it is, formally speaking, irrefutable, just because the data needed to refute it are the very data that are missing. Again, interval statistics leads to qualitatively different results that visually reveal the effect of missing data.

An intermediate case of poor data is the result of what statisticians call censoring. For instance, the measurement of the concentration of rare impurities may be so low as to be below the quantification limits of the laboratory method used to determine its abundance. This does not mean the impurity is not present at all; it only means the lab cannot detect it using its methods. In principle, another lab using more sensitive methods might be able to quantify the concentration. There is a considerable statistical literature on handling censored data like this (Helsel 1990; 2005; Meeker and Escobar 1995). Although many of the available methods are rather sophisticated, the most common approach used in practice is to replace the non-detected datum with a quantity that is one half the detection limit, even though this practice is deprecated by most theorists. Although diagnostics can check the overall fit of models, all of the standard statistical methods for handling censored data make assumptions about the underlying distribution that, insofar as they specify features behind the veil of the censoring, cannot be checked by appeal to data (Manski 2003; Zaffalon 2002; 2005).

When the inferences drawn from data sets with censoring, or missing data, or data of varying quality depend on the specific assumptions made to smooth over these problems, analysts often have genuine concerns about these assumptions. This report explores a different strategy of using intervals to represent this epistemic uncertainty that escapes making such assumptions. Interval statistics elucidates the intrinsic limitations of these kinds of data and allows an analyst to see what the available data are asserting about the underlying population from which they are drawn, without any filter of extra assumptions that cannot, even in principle, be justified empirically.

The underlying motivation behind developing statistics for interval data has close connections to some central ideas in statistics and engineering, including those behind the

fields of robust statistics (Huber 1981) and sensitivity analysis (Saltelli et al. 2001). Like sensitivity analysis, interval statistics asks what can be inferred when data are relaxed from point estimates to reliable ranges. Like robust statistics, it uses weaker but more robust assumptions about the data instead of stronger but more tenuous assumptions employed in conventional analyses. The essential idea behind all of these approaches is that weaker assumptions are more likely to be true. It is sometimes argued that such methods give up analytical power and specificity compared to conventional analyses, but this power and specificity are but pretense if they are based on assumptions that are not warranted. The hope behind interval statistics is to, by this measure, increase the credibility of the conclusions that can be drawn from sample data. All three approaches reject the often pernicious notion seemingly required by conventional analyses of the “omniscient” analyst who somehow knows things about a population under study even though relevant data about it are not available (Huber 1981).

With the increasing availability of powerful computing resources, the sophistication of modeling and simulation has grown enormously over the past decade in all fields, including physics and engineering, but also biology and the environmental sciences. Verification and validation are widely recognized as essential components of modeling and simulation that tether the theorist’s conception to the empirical world. In order to better assess the accuracy of numerical results, the methods for validation must also improve, and the methods of validation depend on the characterization of uncertainties in measurements. The strategic motivation behind this report is to challenge the status quo in metrology to give a more comprehensive accounting of the presence and importance of epistemic uncertainty in some measurements.

*Roadmap to the sections of the report.* The next section of this report reviews why intervals might arise in empirical settings and suggests that data sets containing intervals may be rather commonly encountered in empirical practice. Interval arithmetic and some of its complications are briefly reviewed in section 3 to provide background for the methods developed later in the report. This section also defines several different kinds of interval data sets based on the degree and characteristics of the overlapping among the intervals. A variety of common descriptive statistics for interval data sets are defined in section 4. The section also discusses the development of feasible computer algorithms for evaluating these statistics when simple formulas are not possible. It also gives many numerical examples illustrating the calculation of these statistics for some hypothetical data sets containing intervals. Section 5 introduces the topic of inferential statistics on interval data, and section 6 provides a synoptic review of the computability of the various statistics discussed in the report. Section 7 discusses how sample size and measurement precision interact to affect the uncertainty in a statistic and the nature of the tradeoff between them available to the empiricist. Section 8 reviews the internationally accepted standard methods for handling measurement uncertainty and suggests there might be ways to improve these methods by relaxing some of their assumptions. It suggests that interval statistics might be folded into these standard methods to broaden their generality. Section 8 also includes some numerical examples that compare standard analyses to the interval statistics approach. Finally, section 9 offers some conclusions and suggestions for further research.

There is a great deal of detail and esoterica in several of the sections of this report that is not essential material for most readers. Sections 4, 5 and 6, where the computational algorithms are derived, may be especially burdensome for instance. Of course these calculations generally require a computer for any practical case, and it is unlikely they would ever be done manually by a human. It is entirely reasonable therefore that many readers would elect to skip over some of this text. Rather than destroy the organization of the report by banishing this material to an unstructured series of appendices, we have indicated to the reader where he or she may jump over text without fear of missing something interesting or losing the thread of the argument. These hints are highlighted in gray shading (like this sentence is). Readers who are not particularly interested in some particular statistic or how some calculation will be performed inside software can jump over this material and resume reading at a spot mentioned in the hint.

## 2 Where do interval data come from?

Uncertainty about measurements that is appropriately characterized by intervals is called *incertitude*, and it arises naturally in a variety of circumstances (Ferson et al. 2004b; Osegueda et al. 2002). This section reviews eight sources from which the information is best represented by intervals, including plus-or-minus reports, significant digits, intermittent measurement, non-detects, censoring, data binning, missing data and gross ignorance. At the end of the section, we discuss how intervals differ epistemologically from other kinds of data.

***Plus-or-minus uncertainty reports.*** Engineers and physical scientists are taught to report the plus-or-minus uncertainties associated with the calibration of measuring devices. Such reports often represent epistemic uncertainty in the form of intervals. For instance, a measuring instrument might be characterized by its manufacturer as having a specified reliability such that a measurement taken under prescribed conditions is good to within some margin. An observation using such an instrument would have a form like  $12.64 \pm 0.03$ , which would be associated by rounding to the interval  $[12.61, 12.67]$ . In some cases, the number after the  $\pm$  symbol represents a standard deviation or a standard error, but in other cases, it simply denotes a halfrange. In such cases, there is no statistical or epistemological justification for assuming that the correct value is any closer to the middle of this range than it is to either endpoint (although one might argue that such limits are not absolute in the sense that the true measurement has a zero chance of lying outside of them). For instruments with digital readouts, the genesis of such measurements and their interval-like natures can be apparent. In the case of analog instruments, on the other hand, the act of measurement (mensuration) requires reporting the position of a gauge or needle along a scale on which landmark values are indicated. Many observers report the value associated with that of the nearest landmark as the measurement. After this value has been reported, it is associated with the interval straddling the nearest landmark and extending (usually) half way to the two adjacent landmarks. Some users of analog measurement devices are taught to interpolate to the nearest tenth part between two landmarks. At the limit of the operator's interpolation, whatever it is, there is a kind of uncertainty that can be reasonably and properly represented by an interval. Whether read to the nearest landmark value or interpolated more finely, the observer is responsible for recognizing and reporting the plus-or-minus number associated with the measurement. Note that the origin of the interval in these cases is associated with the *reporting* of the measurement. Even if the mensuration could in principle have been more precise, after this value has been reported, it has the uncertainty as expressed by the observer.

***Significant digits.*** By convention, the number of digits used to express a scalar quantity is used to convey a rough indication of the uncertainty associated with the expression. For example, the unadorned value 12.64 is associated by rounding to the interval  $[12.635, 12.645]$ , without any suggestion about where the actual value is within this range. Because many empirical measurements that are reported lack an explicit statement about their reliability, the convention of interpreting significant digits to implicitly define a

plus-or-minus interval is essential to access these data and use them appropriately for current needs. Although the conventions based on significant digits break down when uncertainty is very large (Denker 2003), interpreting them as defining intervals allows us to make use of an enormous amount of historical and legacy data (and probably some recently collected data) that were collected before the current attention to measurement uncertainty became widespread. The interval implied by the convention of significant digits is, again, associated with the reporting of the measurement as a scalar number with finitely many digits

***Intermittent measurement.*** Some monitoring plans call for periodic or intermittent measurements. A common example of this is the regular inspection of components. If a component is observed to be in good working order at one inspection, but not at the next inspection, when did the component fail? It seems entirely reasonable to conclude that there is a window of time between the last two inspections during which the component failed and that the natural mathematical representation of the failure time is an interval. Such inferences based on temporal observations occur in many domains and cases. For instance, forensic analysis of crime sequencing is often based on synthesis of the testimonies of multiple witnesses that yield an interval window of opportunity.

***Non-detects.*** In chemical quantifications and material purity assessments, there are sometimes “non-detects” in which the laboratory procedure can only say that the concentration of a substance is below a certain amount known as the detection limit. The uncertainty in such a case is in the form of an interval between zero and the detection limit, and there may be no empirical reason to think that the true concentration is more likely to be in any particular part of the interval. For situations in which such uncertainty cannot be neglected in an analysis, it would seem to be essential to have a way to handle these interval uncertainties in calculations.

***Censoring.*** Non-detects described above are sometime said to be left-censored data. There can also be right-censored data. A common example is failure data that includes some components which never failed during the study. For such components, the observed lifetimes should be characterized by an interval from the duration of the study to infinity. It is common for analysts to build a model that tries to project when such components would have failed, but we think it is clear that this goes beyond the *empirical* information at hand. If the conclusion of an analysis depends on extrapolations, it would be important to ascertain what the raw empirical information is telling us.

***Data binning and rounding.*** Another kind of interval measurement arises from binned data, which are sometimes called interval-censored data. Binned data arise when possible values are grouped together to achieve economies in empirical design, to meet limitations on data storage, and as a result of incomplete data transcription. Questionnaire data, for instance, are very commonly binned into a small number of categories (e.g., from questions like “is your age (a) less than 20, (b) between 20 and 50, or (c) over 50?”). Such data are very common in census information of all kinds. Data binning is often the result of restrictions on data collection arising from privacy or security concerns. Binning for the sake of privacy is very common in medical epidemiological information.



For instance, cancer registries may permit access to geographical information about patients only down to the level of census tracts. Information gaps and intentional obscuring of data are becoming more and more common in this time of heightened security. Binned data are usually ordinal in form (although they don't have to be since in principle the intervals could overlap), but the underlying data are not ordinal in nature but rather continuous values that have simply been grouped into categories. Data rounding is a special case of data binning in which the numerical values are grouped together around a nominal value that is an integer or a value with a particular, small number of decimal places. Data rounding is often practiced in manual data transcription and is also common in extremely large data sets such as satellite imagery or continuous-time environmental monitoring data where limitations on computer storage or information transmission rates restrict individual measurements to a few bits or bytes. Rounding has traditionally been modeled with uniform statistical distributions.

**Missing data.** Sometimes data go missing. If the data are missing *at random*, then it might be reasonable to ignore the loss and adjust the sample size down to reflect the number of data values that are available. But it would generally not be reasonable to pretend the data were not planned or collected if they are missing non-randomly for some systematic reason. For instance, if data are more likely to be missing if they represent large magnitudes, failing to account for such missing values would tend to bias the data set. Likewise, in multivariate data sets, if only one dimension out of many is missing, it may not be reasonable to abandon all the measurements that are available just because of one missing value from the set. A practical strategy in these cases is to use intervals to stand for missing data. If a datum was planned but could not be collected or was collected but later lost for some reason, then a vacuous interval  $[-\infty, \infty]$ , or perhaps  $[0, \infty]$  if the value is surely positive, should represent the missing value. The symbol  $\infty$  might represent mathematical infinity, or just the largest possible magnitude a value might take. Even many vacuous intervals in a data set do not necessarily destroy its usefulness. Rather, they allow the epistemic limits embodied in the data to be expressed explicitly, and permit the orderly analysis of the information that is available.

**Gross ignorance.** Sometimes a quantity may not have been studied at all, and the only real information about it comes from theoretical constraints. For instance, concentrations, solubilities, and probabilities and many other kinds of variables have physical limits. These physical limits may be used to circumscribe possible ranges of quantities even when no empirical information about them is available. Typically, these intervals are rather wide, and they are often considered vacuous statements because they represent no empirical information. Nevertheless, mathematical calculations based on them may not always be empty. The uncertainty they induce in computations depends on the uncertainties of other quantities and how all the quantities are combined together in a mathematical expression.

The breadth of the uncertainty captured by the widths of these intervals might be so small as to be negligible. When this is the case, there is no need to bother oneself with any of the methods developed in this report. Indeed, almost all statistical analyses conducted in the past have presumed this to be the case. But the widths of the intervals need not

always be small. In each of the situations outlined above where interval data arise, the uncertainty might be substantial relative to the magnitude of the measurement, or in any case significant for the calculation in which the measurements will be used. When these uncertainties are large enough to matter, it is prudent to take explicit account of them within the analysis.

In this report, an interval characterizing an individual measurement's uncertainty represents an assertion about the measurand's value which is justified by the particular measurement. The interval specifies where the empiricist (or perhaps the instrument) is saying the value is. At the same time it is saying where the value is not. This assertion will be understood to have two mathematical components:

*No concentration:* The interval lacks any concentration of probability or likelihood within the interval, so the actual value is not more likely to be at any one place or another within the interval. But neither is there necessarily a uniform probability distribution over the interval. Instead, the actual value has complete latitude to be anywhere within the interval with probability one. In fact, it might be multiple values within the interval, as would be the case if the quantity is not fixed but varying. The interval represents the fact that we *do not know* the distribution function that should be used to characterize the probability that the quantity is actually one value or another within the interval.

*Full confidence:* The measurement represents an assertion with full confidence that the value is within the specified interval, and not outside it. The assertion is not a statistical confidence statement. The interval is not merely a range in which we have, for instance, 95% confidence the value lies, nor are the intervals such that on average, for instance, 95% of them will enclose their respective true values. That is, it is not a confidence interval or credibility interval. Rather, the interval represents sure bounds on the quantity, to the full degree of confidence that is held by the measurement itself or by the recorder of the measurement.

Both the assumptions of no concentration and full confidence are relaxed in section 8.3. The reasons we start with these assumptions will be evident when we compare the approach developed in this report to the standard approach for evaluating, expressing and propagating measurement uncertainty. Using no-concentration intervals endows our approach with certain desirable features not shared by comparable methods that use uniform or normal (or any other shape) probability distributions (see section 8.1). Using full-confidence intervals greatly simplifies certain calculations, but, of course, a measurand's true value may not actually lie in the given interval just because some measurement asserts it does. The instrument may be improperly calibrated. And, even for well calibrated devices, it is probably not possible to bound all potential measurement errors. For instance, transcription errors and other blunders can occur even in carefully executed measurements. Therefore, the interval bounds may not be entirely accurate even though they are represented as mathematically certain. Nevertheless, such intervals are surely not any more dishonest than using a scalar measurement which implicitly represents itself as having 100% precision (Hajagos 2005, page 96). The whole point of using an interval to model the measurement uncertainty is to begin to acknowledge the intrinsic imprecision in measurement.

### 3 Interval analysis

Section 3.1 briefly reviews the basics of computing with intervals. It is included so the document can be self-contained. Section 3.2 introduces the convention of outward-directed rounding. Section 3.3 reviews the issue of repeated parameters and how it can complicate the calculation of optimal answers in interval analysis problems. Section 3.5 suggests several special cases of data sets containing intervals for which the analyses reviewed in sections 4 and 5 are computationally convenient.

#### 3.1 Interval arithmetic

Interval arithmetic (Young 1931; Dwyer 1951; Moore 1966; 1979; Goos and Hartmanis 1975; Neumaier 1990) is a special case of set arithmetic defined on intervals of the real line. An interval is a closed set of the real line consisting of all the values between an ordered pair of values known as the endpoints of the interval. Mathematically, an interval  $x = [\underline{x}, \bar{x}] = \{X \in \mathbb{R} : \underline{x} \leq X \leq \bar{x}\}$ . In some cases, we might want to define an interval as such a set of the extended real line, denoted by  $\mathbb{R}^* = \mathbb{R} \cup \{-\infty, \infty\}$ , that is, the ordinary set of reals with additional points at positive and negative infinity which are greater and smaller, respectively, than any real number. A real number is a degenerate interval in which the endpoints are coincident. We sometimes call such value ‘scalars’ to emphasize that they are precise and represent only a magnitude and no uncertainty.

Any function defined on real values can be extended to intervals in a straightforward way. In particular, the extension to intervals of a function  $f$  defined on the reals, for intervals  $x, y, \dots, z$ , is

$$f(x, y, \dots, z) = \{f(X, Y, \dots, Z) : X \in x, Y \in y, \dots, Z \in z\}.$$

In other words, when  $f$  is applied to interval arguments, the result is just the set of all possible values that could be obtained from  $f$  by supplying real-valued arguments from the respective intervals. In cases where the result is not an interval because it has holes (i.e., values between the smallest and largest values of the function which are themselves not possible results of the function), it can be made into an interval if desired by taking its convex hull, replacing the set of results by an interval defined by the smallest and largest result values. Evaluating the elementary arithmetic functions of addition, multiplication, etc. does not require actually computing every possible combination of scalars. The formulas for the four basic arithmetic operations are

$$\begin{aligned} x + y &= [\underline{x}, \bar{x}] + [\underline{y}, \bar{y}] = \{X + Y : X \in x, Y \in y\} = [\underline{x} + \underline{y}, \bar{x} + \bar{y}], \\ x - y &= [\underline{x}, \bar{x}] - [\underline{y}, \bar{y}] = [\underline{x} - \bar{y}, \bar{x} - \underline{y}], \\ x \times y &= [\underline{x}, \bar{x}] \times [\underline{y}, \bar{y}] = [\min(\underline{x}\underline{y}, \bar{x}\bar{y}, \underline{x}\bar{y}, \bar{x}\underline{y}), \max(\underline{x}\underline{y}, \bar{x}\bar{y}, \underline{x}\bar{y}, \bar{x}\underline{y})] \text{ and} \\ x \div y &= [\underline{x}, \bar{x}] \div [\underline{y}, \bar{y}] = [\underline{x}, \bar{x}] \times [1/\bar{y}, 1/\underline{y}], \text{ so long as } 0 \notin y. \end{aligned}$$

Excepting any division by zero, the results of these operations always yield compact intervals (closed with no holes). Similar operations such as logarithm, square root, and powers are similarly easy to define so they can extend to interval arguments (Moore 1966; 1979). Keep in mind that we are not saying that every value in an output interval is a possible result of the calculation given the interval bounds on the inputs, only that all values that could be the result are surely in the interval. The results of these interval operations are guaranteed to enclose all the values that could possibly be results of the calculations (Moore 1966; Neumaier 1990). We say that these results are *rigorous* bounds.

### **3.2 Outward-directed rounding**

In interval analysis, whenever numbers must be rounded in calculations or for numeric display, outward-directed rounding is conventionally employed to guarantee a reported result will contain all the possible values. Outward-directed rounding ensures the rigorousness of the interval calculation is not lost. For example, the theoretical interval  $[2/3, 10/3]$  should be rounded to something like  $[0.666, 3.334]$ , even though the quotients  $2/3$  and  $10/3$  considered as scalar values would ordinarily be rounded to 0.667 and 3.333. If the endpoints of the interval had instead been rounded to the nearest three-place scalar values, the displayed interval would fail to be rigorous because it would be excluding some of the possible values in the theoretical interval, namely the values in the ranges  $[0.66\bar{6}, 0.667]$  and  $[3.333, 3.33\bar{3}]$ . In computer science applications of interval analysis (Alefeld and Hertzberger 1983), this convention is used for internal calculations when the bytes devoted to a numeric quantity do not permit a perfect representation of the quantity. When there are many sequential calculations, the errors from ordinary rounding to the nearest representable scalar can accumulate and lead to a serious diminution of accuracy. In this report, however, outward-directed rounding will primarily be an issue associated with the output display of numerical results. The convention can sometimes lead to answers that initially seem surprising. For instance, when the output precision is set to display only three significant digits, the theoretical interval  $[0, 1.0000000000000001]$  would be rounded to  $[0, 1.01]$ , which is the best interval expressible in terms of numbers with three significant digits that encloses the theoretical interval.

### **3.3 Repeated parameter problem**

In many problems, interval arithmetic based on the rules described in section 3.1 can be used to obtain results that are both rigorous and best possible. However, when an uncertain number appears more than once in a mathematical expression, the straightforward sequential application of the rules of interval arithmetic may yield results that are wider than they should be. Blithely applying the rules this way is sometimes called “naive” interval analysis because it risks losing its optimality. The result is still rigorous in that it is sure to enclose the true range, but it fails to be best-possible if it is wider than it needs to be to do so. The reason for this loss of optimality is basically that the uncertainty in the repeated parameter has entered into the calculation more than once.

The appearance of repeated parameters in expressions is a well-known problem with interval arithmetic and, indeed, with all uncertainty calculi (e.g., Moore 1966; Manes 1982; Hailperin 1986; Ferson 1996). In interval analysis, the problem is less severe than in other calculi because its methods always fail safe in the sense that they cannot underestimate the uncertainty. This is not the case in probabilistic arithmetic (Springer 1979), where introducing a repeated parameter says that variable is independent of itself, which is nonsensical and can underestimate tail probabilities. Mathematical expressions without repeated parameters, sometimes called single-use expressions (or SUEs), are guaranteed to yield optimal results under interval analysis, i.e. rigorous bounds that are as tight as possible given the inputs (Moore 1966). But repetition of uncertain parameters does not automatically doom the result of naive interval analysis to suboptimality. For instance,  $x + x + x$  will be as tight as possible despite the repetition of the interval  $x$  if the values in  $x$  have constant sign.

There is no general algorithm that can calculate optimal bounds for functions with interval parameters in polynomial-time (Kreinovich et al. 1997). This means that, when there are many intervals involved in a calculation—such as is typical in interval statistics—the computational effort required to obtain the optimal bounds exactly can be prohibitively large. However, one may be able to devise heuristic approaches to handle different subclasses of problems with more modest computational effort. In interval analysis many strategies have been developed to deal with the excess width problem. One of the most direct methods is to algebraically manipulate the expression to reduce the occurrences of the repeated parameters, such as replacing the original expression  $x + x + x$  with the equivalent expression  $3x$ , or completing the square (Jaulin et al. 2001) by which  $x + x^2$  is re-expressed as  $(x + \frac{1}{2})^2 - \frac{1}{4}$ , which looks a bit more complex but has only one instance of the uncertain parameter  $x$  and therefore can be easily computed with the naive methods.

### ***3.4 Two interpretations of an interval***

There are two theories about what an interval represents. The first holds that an interval is simply a collection of all the real numbers—what we have been calling scalars—that lie between two endpoints. In this interpretation, the elements of an interval  $[A, B]$  could be depicted on a plot of cumulative probability versus the quantity as a collection of vertical spikes\* arranged at every possible scalar value from  $A$  to  $B$ . Such a display is shown in the left graph in Figure 1. There are infinitely many of these spikes of course, and the figure can only show some of them, but they form a simple set that might be described as a geometric comb for which the spikes are the teeth.

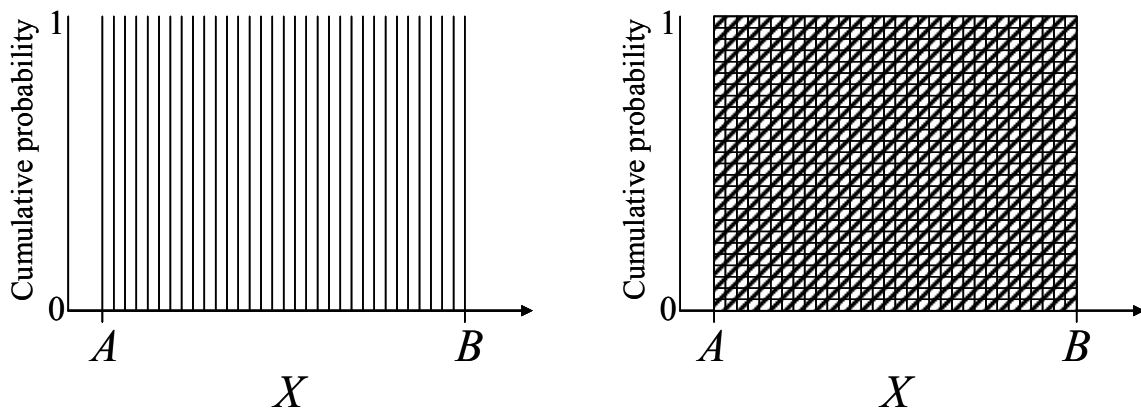
The second theory about intervals holds that they represent a much larger variety of uncertainty. Under this interpretation, an interval  $[A, B]$  represents the set of all

---

\*Such a spike is sometimes called the delta distribution for the value. It is a degenerate probability distribution that lacks variability. Each spike actually has a ‘left tail’ at probability level zero extending from negative infinity to the foot of the spike, and an analogous ‘right tail’ at probability level one extending from the top of the spike to positive infinity. These tails are not shown in the figure.

probability distribution functions  $F$  defined on the real numbers such that  $F(X) = 0$  for all values  $X < A$ , and  $F(X) = 1$  for all values  $X > B$ . The graph on the right side of Figure 1 tries to depict this case, although it is even harder because the set of distributions includes all of the spikes, plus a huge variety of other distributions. For instance, the collection includes all possible horizontal lines at any probability level between zero and one from  $A$  to  $B$ . These correspond to scaled Bernoulli distributions that have some mass at  $A$  and the rest at  $B$ . It also includes all increasing straight lines from the left or bottom edges of the box to the top or right edge. Such lines represent distributions that smoothly distribute some of their mass across the range from  $A$  to  $B$ , and the rest of their mass at one or both endpoints (depending on whether the line intersects the interiors of the left or right edge of the box). In fact, the collection of distributions includes absolutely *any* nondecreasing curve from the left or bottom edges of the box to the top or right edge.

We will distinguish between the two theories of intervals when we need to by calling the first the ‘spike interpretation of intervals’ and the second the ‘box interpretation of intervals’. When the term ‘interval’ would be ambiguous under the two possible interpretations, we might also refer to the collection of scalar values under the spike interpretation as a ‘scalar range’ and the collection of arbitrary distributions as a ‘p-box’. Both of the interpretations about what an interval is can be found in the literature on interval analysis, although the difference between them seems to have been important only rarely so the distinction has not been emphasized. This report will show that the two interpretations lead to different theories about the statistics of data sets containing intervals, both of which are meaningful and have practical uses. Although there are many situations and calculations for which it won’t make any difference which interpretation is used, the exceptions are interesting and important. Because a scalar range is a proper subset of the corresponding box, any statistic computed for a data set of intervals under the spike interpretation will be a subset of the statistic computed for the data set under the box interpretation. In section 4.5.1, for instance, we shall see that this fact comes in handy to develop a convenient conservative algorithm for the variance.



**Figure 1. Two interpretations about what an interval represents.**

### 3.5 Kinds of interval data sets

We shall see in this report, especially in section 4, that the ease of computing several statistics depends strongly on the nature of the data set involving interval values. Narrow intervals make some calculations easy and some hard. Broadly overlapping intervals make some calculations easy and complicate other computations. The narrowness and scatteredness of the intervals play roles, as does the regularity of the widths of the intervals. This section introduces a variety of kinds of interval data sets that we might reasonably expect to encounter.

**No intersections.** Some of the computational problems considered in the following sections become much easier when no intervals intersect one another. What kinds of interval data have no intersections? This should happen, of course, if the measurements of generic real-valued quantities are sufficiently precise that the corresponding intervals become so narrow that they avoid intersecting each other. We might also have no intersections if the dispersion of the underlying distribution is very wide relative to the imprecision of individual measurements. Note that a no-intersections data set is a relaxation of statistical models in use today that presume that—or at least act as if—measurements are infinitely precise. It is relaxed because the measurements may still express incertitude so long as it is not so broad to lead to intersections among the intervals.

**Same precision.** In some situations, the uncertainties associated with each measurement are all the same. This can happen, for instance, when the uncertainties arise from the readouts on a digital measuring device that displays a fixed number of decimal places, or when measurements are made on a single scale with constantly spaced landmarks. In such situations all measurements are performed with the same precision. In mathematical terms, this would result in all intervals  $[\underline{x}_i, \bar{x}_i]$  having the same width so that  $\bar{x}_i - \underline{x}_i$  is a constant  $\Delta$  for all  $i$ . In fact, the algorithms described in this report that work for collections of intervals having the same precision also work if the data sets include scalars as well as intervals. Consequently, we can define a same-precision data set to be one whose *nondegenerate* intervals all have the same width. Like the no-intersection data sets defined above, same-precision data sets generalize traditional data sets composed entirely of scalar point values, but this generalization allows arbitrarily large overlap among the intervals.

**Binned data.** In many situations, data are binned into adjacent intervals. For instance, data about individual humans is often collected under restrictions intended to protect privacy rights. One of the standard restrictions is to collect only data coarsely binned into non-overlapping intervals. Similar data sets also arise in checklist and condition scoring surveys in engineering, biology and social science. In such data sets, intervals either coincide (if the values belong to the same range) or they are different, in which case they are disjoint or can only touch only at a common endpoint. A data set having the property that any two nondegenerate intervals either coincide or intersect in at most one point is called a binned data set. A binned data set can also contain arbitrary scalar values.

**Detect/non-detect.** If the only source of interval uncertainty is detection limits, i.e., if every measurement result is either a detect represented by an exact scalar value or a non-detect represented by an interval  $[0, DL_i]$  for some real number  $DL_i$  (with possibly different detection limits for different sensors), then we will call the collection a detect/non-detect data set. Data sets fitting this definition commonly arise in chemical and materials characterization laboratories. An algorithm that works for the binned-data case will also work for the detect/non-detect case if all sensors have the same detection limit  $DL$ .

**Few intersections.** If the no-intersections case is relaxed so that the intervals become wider, there may be some intersections among them, but we might expect to see few intervals intersecting. We can quantify this idea with an integer limit  $K$  on the number of overlapping intervals in the data set. If a data set is said to have “few intersections of order  $K$ ”, no set of  $K$  intervals has a common intersection. When  $K = 2$ , this corresponds to the previous no-intersections case.

**No nesting.** An interval data set is said to have no nesting if no nondegenerate interval in the data set is a subset of the interior of another. Data sets collected using a single measuring instrument, or perhaps using different measuring instruments of the same type, are likely to have no nesting. The algorithms that have been developed for this kind of interval data are insensitive to the presence of point values, so, in practice we can allow any scalar values that may be in a data set to be nested within intervals and still call it a no-nesting data set. A data set consisting of  $N$  intervals  $[\underline{x}_i, \bar{x}_i]$ ,  $i = 1, \dots, N$ , has no nesting, or has the no-nesting property, if  $\underline{x}_i \neq \bar{x}_i$  always implies  $[\underline{x}_i, \bar{x}_i] \not\subset ]\underline{x}_j, \bar{x}_j[$  for all  $i = 1, \dots, N$ , and any  $j = 1, \dots, N$ , and where the reverse brackets denote the open interval excluding the two endpoints. The no-nesting case generalizes all of the cases already mentioned. No-intersections data, same-precision data, binned data, and detect/non-detect data also satisfy the no-nesting property, so any algorithms that work for the no-nesting case will work for these cases as well.

**Thin.** A relaxation of the condition of no nesting can define a different kind of narrowness among the intervals in a data set. If the intervals exhibit the no-nesting property when they are hypothetically narrowed by symmetrically dividing their widths by the sample size  $N$ , then the data set is said to be thin. Symbolically, thinness in an interval data set means that for all intervals  $i \neq j$

$$\left[ \tilde{x}_i - \frac{\Delta_i}{N}, \tilde{x}_i + \frac{\Delta_i}{N} \right] \not\subset \left[ \tilde{x}_j - \frac{\Delta_j}{N}, \tilde{x}_j + \frac{\Delta_j}{N} \right]$$

where  $\tilde{x}$  denotes the midpoint of an interval  $(\underline{x} + \bar{x})/2$ , and  $\Delta$  is its halfwidth  $(\bar{x} - \underline{x})/2$ . This condition is implied by

$$\left| \frac{\bar{x}_i - \underline{x}_i}{2N} - \frac{\bar{x}_j - \underline{x}_j}{2N} \right| \leq \left| \frac{\underline{x}_i + \bar{x}_i}{2} - \frac{\underline{x}_j + \bar{x}_j}{2} \right|$$



which, in turn, can be reduced to checking that

$$|\bar{x}_i - \underline{x}_i - \bar{x}_j + \underline{x}_j| \leq N |\underline{x}_i + \bar{x}_i - \underline{x}_j - \bar{x}_j|$$

for all  $i \neq j$ . This condition is a generalization of the no-nesting property, which means that algorithms that work for thin data will also work with no-nesting data and all of its special cases as well. Thin data can display considerable overlap among the intervals.

**Arrangeable.** The no-nesting condition can also be relaxed in a different way. A natural next case is when the interval measurements can be divided into subgroups, each of which has the no-nesting property. This situation might arise from the use of a few different measuring instruments.

**Scattered.** Another relaxation of the condition of no intersections can define a different kind of scatteredness among the intervals in a data set. If the intervals exhibit no or few intersections when they are hypothetically narrowed by symmetrically dividing their widths by the sample size  $N$ , then the data set is said to be scattered. In the same way that thin data generalize no-nesting data, scattered data generalize no-intersections data. Symbolically, an interval data set  $x_i, i = 1, \dots, N$ , is called scattered if there are no more than  $K$  intervals among the narrowed intervals  $y_i$  that have a common intersection, where

$$y_i = \left[ \tilde{x}_i - \gamma \frac{\Delta_i}{N}, \tilde{x}_i + \gamma \frac{\Delta_i}{N} \right] = \left[ \underline{x}_i + \bar{x}_i - \gamma \frac{(\bar{x}_i - \underline{x}_i)}{N}, \underline{x}_i + \bar{x}_i + \gamma \frac{(\bar{x}_i - \underline{x}_i)}{N} \right] / 2,$$

where  $i = 1, \dots, N$ ,  $\gamma$  is a constant, and  $\tilde{x}$  and  $\Delta$  denote the midpoint and halfwidth of an interval. These narrowed intervals have the same midpoints as the original data  $x_i$ , but they can have much smaller widths, depending on  $\gamma$  and the sample size  $N$ .

**General interval data.** Finally, any data set that does not fit in one of the previous classes will be considered a part of the general class of interval data.

Figure 2 gives examples of each of these kinds of data sets defined above. In each kind of data set, the intervals are displayed as horizontal lines whose location on the  $X$  axis is given by their endpoints. They are displaced vertically so they are easy to distinguish, but these vertical positions represent only the order in which the data were collected and are otherwise statistically meaningless.

Notice that the data set in the second panel of the figure exemplifying same-precision data has a rather large imprecision, resulting in substantial overlap among the intervals. The binned data in the third panel also exhibit a great deal of overlap, but the overlap comes in only two forms; either intervals overlap completely or they only touch at their ends. In the example detect/non-detect data, there are three non-detects and five scalar values (which are depicted as stubby lines for visual clarity). The data set illustrating the few-intersections case shows that there can be many and broad intersections among the

intervals; there just cannot be too many intervals involved in any particular intersection of  $X$ -values. The example data set depicted in the fifth panel of the figure has no more than two intervals with a common intersection, so it represents a few-intersections data set of order  $K=3$ . The sixth panel of data in the figure illustrates the no-nesting property, but so do the first four panels of data as they are special cases of no nesting.

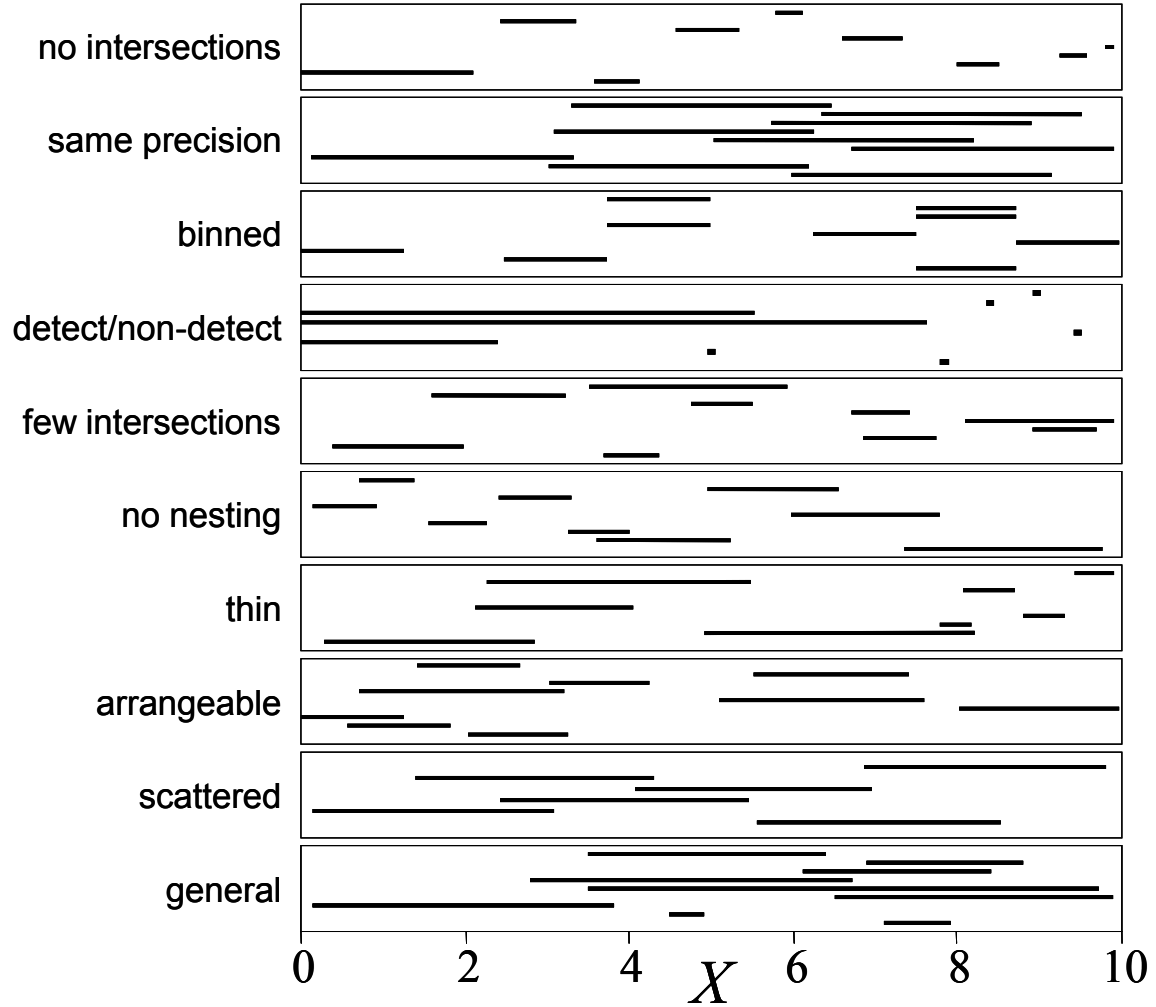


Figure 2. Examples of the various kinds of interval data sets.

It may often be possible for an analyst to discern by visual or graphical inspection whether or not a data set is one of the first six kinds of interval data. This will typically not be possible to do for the other categories. Although the figure also depicts example data sets that are thin, arrangeable, and scattered, it is hard to capture these properties in such examples. A huge variety of data sets are thin, for instance, and it is rather hard to develop any intuition about what the graphs for thin, or arrangeable or scattered, data sets should look like. These properties usually must be checked by the computer. At the bottom of the figure is a general data set that has none of the nine properties defined in this section.

Figure 3 shows the relationships among the kinds of interval data sets. The arrows point from general to special cases.

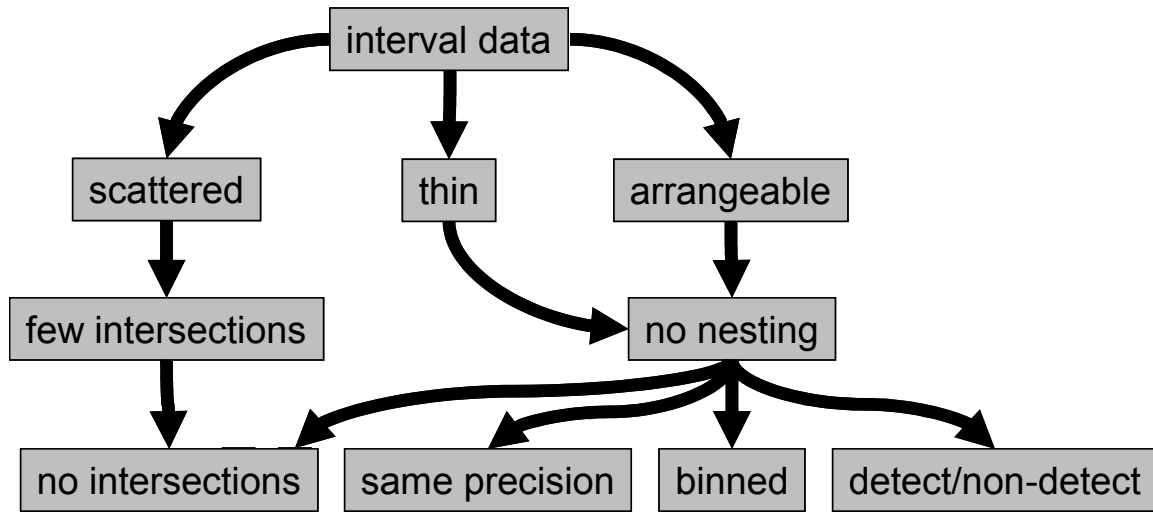


Figure 3. Relationships among the different kinds of interval data sets.

## 4 Descriptive statistics for interval data

This section reviews a variety of basic descriptive statistics such as means, variances and confidence intervals, for interval data sets, and the formulas and algorithms for computing these statistics. Section 5 introduces the subject of inferential statistics which is used to make statistical conclusions or inferences from interval data via statistical tests such as regressions, and section 6 summarizes the computational feasibility of the various algorithms when sample sizes become large.

Descriptive statistics is concerned with describing or characterizing data sets. This description is a summarization of the data themselves and is usually accomplished by computing quantities called statistics that characterize some important feature of the data set. For instance, there are statistics that characterize the central tendency of a collection of data (such as the mean and median), and there are statistics that characterize the dispersion of the data (such as the variance and interquartile range). Some of the algorithms for computing these statistics for interval data sets are straightforward extensions of analogous methods used for scalar data (as in the case of the mean). But some of the algorithms are considerably more complicated than their analogs used for scalar data (as in the case of the variance). In most cases, the additional complexity arises from the repeated parameter issue discussed in section 3.3.

In the discussions below, we will use the following notation to refer to interval data sets, and to two example data sets. We will suppose the empirical focus is on some quantity  $X$  for which we have a collection of  $N$  measurements  $x_i$ ,  $i = 1, \dots, N$ , where  $x_i = [\underline{x}_i, \bar{x}_i]$  are intervals, i.e.,  $\underline{x}_i \leq \bar{x}_i$ . Each of these intervals represents bounds on the empirical uncertainty for the corresponding measurement. The data may include point estimates, i.e., scalar values, as special cases for which the lower and upper endpoints are coincident.

This section gives numerical examples illustrating the computation of the various statistics as they are introduced. These examples will make use of two small hypothetical data sets which we now describe. In the first data set, which we will call ‘skinny’, the intervals are so narrow as to not overlap each other. In the second data set, ‘puffy’, there are half again as many measurements, but the intervals are generally wider and exhibit a lot of overlap. These are the data values:

Skinny data	Puffy data
[1.00, 1.52]	[3.5, 6.4]
[2.68, 2.98]	[6.9, 8.8]
[7.52, 7.67]	[6.1, 8.4]
[7.73, 8.35]	[2.8, 6.7]
[9.44, 9.99]	[3.5, 9.7]
[3.66, 4.58]	[6.5, 9.9]
	[0.15, 3.8]
	[4.5, 4.9]
	[7.1, 7.9]

These two data sets are depicted in Figure 4 and Figure 5. The skinny data set is a model of a traditional data set in which the imprecision represented by the intervals is very narrow. This is a no-intersections data set as described in section 3.5. The puffy data set is a general interval data set in that it does meet any of the definitions of special cases of interval data described in section 3.5.

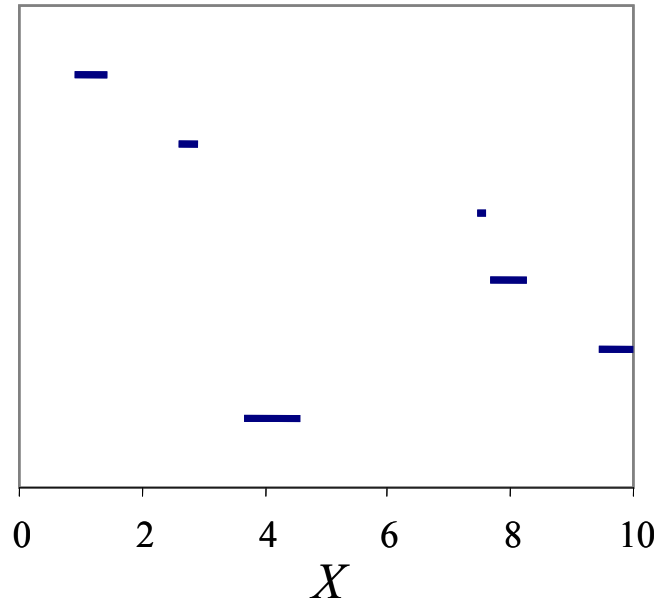


Figure 4. Hypothetical interval data set ‘skinny’ consisting of 6 measurements for a quantity  $X$ . (Vertical position of the intervals is not meaningful.)

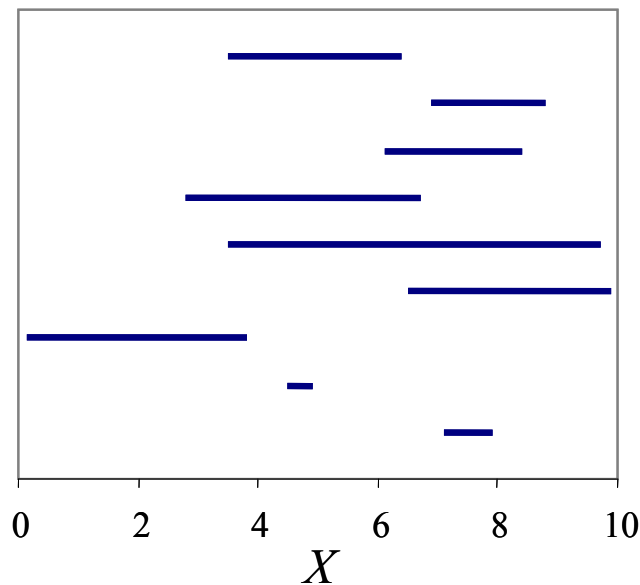


Figure 5. Hypothetical interval data set ‘puffy’ consisting of 9 measurements for a quantity  $X$ . (Vertical position of the intervals is not meaningful.)

## 4.1 Distribution

This section describes how the empirical distribution function for scalar data can be generalized for interval data. Subsections 4.1.2 and 4.1.3 detail the calculation of confidence limits on the distribution and the use of weighted data. (The subject of fitting distributions to interval data sets with assumptions about the shape or family of the distribution is addressed later in section 4.9.)

A distribution function for a data set, which is sometimes called an empirical distribution function or EDF, summarizes the data set in a graphical form. It is a function from the  $X$ -axis (i.e., whatever the measured parameter is) to a probability scale of  $[0, 1]$ . The distribution of a data set of precise scalar values is constructed as an increasing step function with a constant vertical step size is  $1/N$  where  $N$  is the sample size. The locations of the steps correspond to the values of the scalar data points. Such a distribution is denoted by the function  $S_N(X)$  which is the fraction of data values in the data set that are at or below the magnitude  $X$ . A distribution preserves the statistical information in the data set about its central tendency or location, its dispersion or scatter, and, in fact, all other statistical features of the distribution. The only information in the original data set that is not in the distribution is the order in which the values are given. Because most analyses assume that the data are selected at random, this information is not meaningful anyway.

The analog of a precise empirical distribution function for a data set that contains interval values is a p-box (Ferson et al. 2003). The p-box is the set of all (precise) distribution functions that could arise from any possible configuration of scalar values from within the respective intervals that make up the data. Suppose, for instance, that the data set consists of only the two intervals  $[1,3]$  and  $[2,4]$ . The distribution for these two data would be the collection of all two-point distribution functions having one foot, so to speak, in the first interval and one in the second interval. There are an infinite number of such pairs of points. The table below lists a few of them.

	<b>1</b>	<b>1.1</b>	<b>1.2</b>	<b>1.3</b>	<b>...</b>	<b>2.8</b>	<b>2.9</b>	<b>3</b>
<b>2</b>	{1.0,2.0}	{1.1,2.0}	{1.2,2.0}	{1.3,2.0}		{2.8,2.0}	{2.9,2.0}	{3.0,2.0}
<b>2.1</b>	{1.0,2.1}	{1.1,2.1}	{1.2,2.1}	{1.3,2.1}		{2.8,2.1}	{2.9,2.1}	{3.0,2.1}
<b>2.2</b>	{1.0,2.2}	{1.1,2.2}	{1.2,2.2}	{1.3,2.2}		{2.8,2.2}	{2.9,2.2}	{3.0,2.2}
<b>2.3</b>	{1.0,2.3}	{1.1,2.3}	{1.2,2.3}	{1.3,2.3}		{2.8,2.3}	{2.9,2.3}	{3.0,2.3}
<b>2.4</b>	{1.0,2.4}	{1.1,2.4}	{1.2,2.4}	{1.3,2.4}		{2.8,2.4}	{2.9,2.4}	{3.0,2.4}
<b>2.5</b>	{1.0,2.5}	{1.1,2.5}	{1.2,2.5}	{1.3,2.5}		{2.8,2.5}	{2.9,2.5}	{3.0,2.5}
<b>⋮</b>					<b>⋮</b>			
<b>3.8</b>	{1.0,3.8}	{1.1,3.8}	{1.2,3.8}	{1.3,3.8}		{2.8,3.8}	{2.9,3.8}	{3.0,3.8}
<b>3.9</b>	{1.0,3.9}	{1.1,3.9}	{1.2,3.9}	{1.3,3.9}		{2.8,3.9}	{2.9,3.9}	{3.0,3.9}
<b>4</b>	{1.0,4.0}	{1.1,4.0}	{1.2,4.0}	{1.3,4.0}		{2.8,4.0}	{2.9,4.0}	{3.0,4.0}

The top margin of the matrix (in bold type) gives the value from the interval  $[1,3]$ , and the left margin of the matrix gives the value from the interval  $[2,4]$ . The first pair in the upper, left corner of the matrix consists of the left endpoint from interval  $[1,3]$  and the

left endpoint from interval  $[2,4]$ . The table implies  $21 \times 21$  pairs (and lists only  $9 \times 7$  explicitly), but clearly there are infinitely many other combinations of values from the two intervals, because there are infinitely many distinct values in  $[1,3]$  and infinitely many in  $[2,4]$ . The set of all of these pairs  $\{(a,b) : a \in [1,3], b \in [2,4]\}$  is the Cartesian product space  $[1,3] \times [2,4]$ . Each such pair is a possible distribution. In general, when there are many intervals  $x_1, x_2, \dots, x_N$  in an interval data set, they likewise correspond to a Cartesian space  $[\underline{x}_1, \bar{x}_1] \times [\underline{x}_2, \bar{x}_2] \times \dots \times [\underline{x}_N, \bar{x}_N]$ . If we presume that the  $N$  measurands had  $N$  precise values—and only measurement uncertainty prevented the empiricist from identifying them precisely—then the true underlying data set, which our interval set bounds, is some point in this Cartesian space, which is an  $N$ -dimensional hypervolume. We sometimes refer to a point in this space as a ‘configuration’ of scalar values within their respective intervals. The distribution for the interval data set is the collection of all distributions for any configuration of scalar values within their respective intervals. Likewise, as we shall see in the rest of this report, any statistic for an interval data set is the set of possible values for the same statistic over all possible configurations of scalar values within their respective data intervals, or, often more simply, bounds on that set.

The bounds for the distribution of an interval data set can be computed very simply. These bounds are the p-box formed by enveloping the (precise) distribution associated with the left endpoints of all the intervals  $\{\underline{x}_1, \underline{x}_2, \dots, \underline{x}_N\}$  with the distribution associated with the right endpoints  $\{\bar{x}_1, \bar{x}_2, \dots, \bar{x}_N\}$ . The left and right sides of the resulting p-box are

$$[\underline{B}(X), \bar{B}(X)] = [S_{LN}(X), S_{RN}(X)]$$

where  $S_{LN}(X)$  is the fraction of the left endpoints in the data set that are at or below the magnitude  $X$ , and  $S_{RN}(X)$  is the fraction of right endpoints in the data that are at or below  $X$ . This p-box bounds all of the distributions associated with the infinitely many configurations in the  $N$ -dimensional Cartesian space,

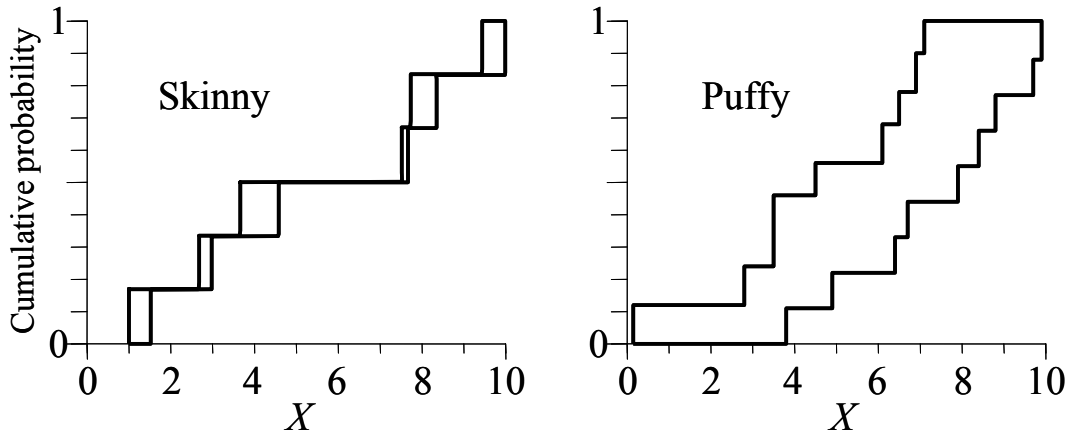
$$\underline{B}(X) \geq S_N(X) \geq \bar{B}(X)$$

where  $S_N(X)$  is the distribution function for any configuration of scalar values. There is a difference between the collection of distributions for each point in the  $N$ -dimensional hypervolume represented by interval data and the p-box that bounds that collection. When we refer to ‘the distribution’ for an interval data set, we will schizophrenically sometimes be thinking of the *collection* and sometimes of its bounding *p-box*. Either context or explicit definition can always make it clear which is intended. Using the p-box as the distribution of an interval data set is, of course, vastly simpler than trying to use the infinite collection of distributions associated with all possible configurations of scalars within these intervals.

Like the case for scalar data, the distribution function for interval data forgets the original ordering of the data set. In the interval case, however, the distribution as a p-box also forgets which endpoints are associated with which endpoints. Thus, if we use the

bounding p-box to represent a generalized distribution, the distribution associated with a data set consisting of the two intervals  $[1,3]$  and  $[2,4]$  cannot be distinguished from the distribution based on the intervals  $[1,4]$  and  $[2,3]$ . Both have left endpoints at 1 and 2, and both have right endpoints at 3 and 4. This information loss implies that one cannot generally fully reconstruct from a p-box distribution of interval data the original values on which it was based. The distribution of interval data, like its analog for scalar data, characterizes the probability that a randomly selected datum will have a value less than or equal to any given magnitude. Because of the incertitude in the data set, this probability will in general be an interval, rather than a scalar value.

*Numerical examples.* The p-box representing the distribution of the skinny data set is formed by enveloping the distributions for the scalar data sets  $\{1.00, 2.68, 3.66, 7.52, 7.73, 9.44\}$  and  $\{1.52, 2.98, 4.58, 7.67, 8.35, 9.99\}$ . Likewise, the p-box for the puffy data set is the pair of distributions for  $\{0.15, 2.8, 3.5, 3.5, 4.5, 6.1, 6.5, 6.9, 7.1\}$  and  $\{3.8, 4.9, 6.4, 6.7, 7.9, 8.4, 8.8, 9.7, 9.9\}$ . The resulting p-boxes are depicted in Figure 6.



**Figure 6.** Empirical distributions for the skinny and puffy data sets containing interval uncertainties.

The motivation and formulation for this analog of the distribution for an interval data set have relied on the spike interpretation of intervals as discussed on page 28. We could have instead used the box interpretation of intervals. Under the box interpretation, the intervals represent all the distribution functions having support on that range, and the aggregation into an overall distribution of the various intervals in the data set is a mixture operation (Ferson et al. 2003). The resulting collection of possible distributions will not only include two-point mass distributions, but infinitely many other distributions as well. For the earlier example with only two intervals  $[1,3]$  and  $[2,4]$ , this collection is simply  $\{H : H(X) = F(X)/2 + G(X)/2, X \in \mathbb{R}, F \in \mathcal{F}_{[1,3]}, G \in \mathcal{F}_{[2,4]}\}$ , where

$$\mathcal{F}_{[a,b]} = \{F : X < a \rightarrow F(X) = 0, b < X \rightarrow F(X) = 1, X < Y \rightarrow F(X) \leq F(Y), 0 \leq F(X) \leq 1\}$$

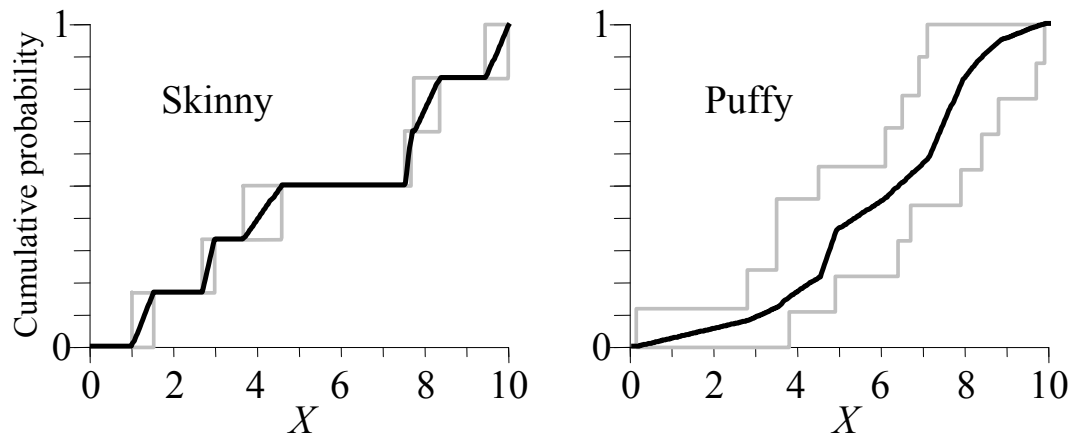
is the set of distribution functions over a specified support  $[a,b]$ . Despite the vast increase in the cardinality of the collection, the outer bounds on this set of distributions



that could arise as a mixture of distributions from each input interval nevertheless remains exactly the same as under the spike interpretation.

#### 4.1.1.1 Equiprobability model of intervals

Several analysts have suggested an alternative treatment for interval data based on modeling each interval as a uniform distribution (e.g., Bertrand and Groupil 2000; Billard and Diday 2000; Billard and Diday n.d.; Bock and Diday 2000; cf. Gioia and Lauro 2005). Bertrand and Groupil (2000) call this the “equidistribution hypothesis” and it represents the idea that each possible value in an interval is equally likely. Billard and Diday (n.d.) survey some of the descriptive statistics that can be computed from interval data under this model. An empirical distribution under this model would be formed as an equal-weight mixture of the uniform distributions representing the respective intervals. Figure 7 displays such empirical distributions for the skinny and puffy data as black curves, which are superimposed over the gray p-boxes resulting from using the interval statistics approach described above. As can be seen in the graphs, this approach is a way to split the difference, as it were, in estimating the empirical distribution. Because the black curves are ordinary, precise distributions, their summary statistics are fairly easy to compute.



**Figure 7. Empirical distributions for the skinny and puffy data sets under the equidistribution hypothesis (black) superimposed on their interval-statistics analogs (gray).**

Although this approach of modeling intervals as little uniform distributions may be useful in some analyses, it is incompatible with the perspectives on incertitude adopted in this report. The idea that ignorance should be represented by equiprobability is a central idea in probability theory, dating back at least to Laplace (1820) himself, but it is exactly this idea that interval statistics as conceived here is designed to relax. Bertrand and Groupil (2000) acknowledge that, even as the sample size grows to infinity, the true limiting distribution of the underlying population is only approximated by any approach based on the equidistribution hypothesis. The reason, of course, is that the uniform distributions may not be accurate reflections of probability within the respective intervals. The approach suggested in this report will provide ways to obtain conclusions that do not depend on whether probability mass is uniform or not.

You may elect to skip the following two minor subsections and resume reading with section 4.2 on page 36.

### 4.1.2 Confidence limits on the distribution

Kolmogorov (1941) described distribution-free confidence limits on a distribution function characterized by random samples, that is, by values selected independently from a common continuous distribution. The confidence limits depend on the value of a special statistic given by Smirnov (1939), thus the bounds are often called the Kolmogorov-Smirnov confidence limits (Feller 1948; Miller 1956; Ferson et al. 2003, section 3.5.4). The bounds are computed by vertically expanding the upper and lower bounds of the distribution by a constant that depends on the sample size and the level of confidence desired. The left and right sides of the resulting p-box are

$$[\underline{B}(X), \bar{B}(X)] = [\min(1, S_{LN}(X) + D_{\max}(\alpha, N)), \max(0, S_{RN}(X) - D_{\max}(\alpha, N))] ]$$

where, again,  $S_{LN}(X)$  is the fraction of the left endpoints of the values in the data set that are at or below the magnitude  $X$ ,  $S_{RN}(X)$  is the analogous fraction of right endpoints in the data that are at or below  $X$ , and  $D_{\max}$  is the Smirnov statistic that depends on both the confidence level and the sample size. The result has a classical frequentist (Neyman-Pearson) interpretation. The statistical statement associated with these confidence limits is that  $100(1-\alpha)\%$  of the time such bounds are constructed (from independent data selected randomly from a common, continuous distribution), they will totally enclose the true distribution function from which the samples were drawn. Thus, to compute 95% confidence limits,  $D_{\max}(0.05, N) \approx 1.36/\sqrt{N}$ , so long as  $N$  is larger than about 50. For smaller values of  $N$ , the value of  $D$  must be read from a statistical table (e.g., Miller 1956; Crow et al. 1960, Table 11, page 248). The Kolmogorov-Smirnov confidence limits are distribution-free, which means that they do not depend on any particular knowledge about shape of the underlying distribution. They only require it be continuous and that the samples characterizing it be independent and drawn randomly from this distribution.

Theoretically, the left tail of the left bound  $\underline{B}$  is  $D_{\max}(\alpha, N)$ , and extends to negative infinity without ever declining. In practical use, however, we may know that a variable must be positive and, if so, we can use that information to truncate this left limit to zero for negative numbers. The right tail of the right bound  $\bar{B}$  likewise extends to positive infinity at the probability level  $1 - D_{\max}(\alpha, N)$ , but knowledge of any constraints on high values of the variable could be used to truncate this tail too.

*Numerical examples.* The 95% confidence limits on the empirical distributions for the skinny and puffy data sets are shown in Figure 8 as solid black lines. The empirical distributions for the two data sets (that were previously shown in Figure 6) are echoed here as dotted gray lines. For the skinny data set, which contains 6 samples,  $D_{\max}(0.05, 6) = 0.525$ , and for the puffy data set with 9 samples,  $D_{\max}(0.05, 9) = 0.434$ . Because these data sets have so few samples, the confidence limits are quite wide. In these cases, the breadth of the limits is almost entirely due to the paucity of samples. However, if there had been many samples of wide intervals, this result could be the other way around. Note that the upper bounds on the left tails and the lower bounds on the right tails of

these distributions extended in either direction to infinity without approaching zero or one. This has nothing to do with the interval uncertainty in the data sets; it is a feature of the Kolmogorov-Smirnov confidence limit and is seen in traditional applications on scalar data sets.

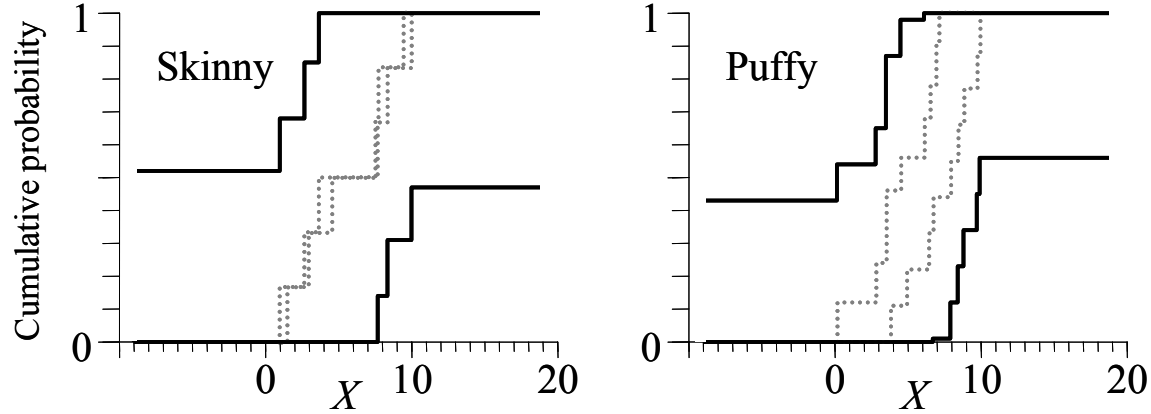


Figure 8. 95% confidence limits (black) on the distribution functions corresponding to the empirical distribution functions (dotted gray).

### 4.1.3 Weighted data

Many data sets are not simply observations of an  $X$  variable, but sometimes include weights associated with each such observation. This might happen, for instance, if data collection was not strictly random, but stratified-random. In such a case, the probability of being in a stratum would be used as the weight associated with the value observed (randomly) from that stratum. If data potentially including incertitude are accompanied by scalar weights, the distribution described in section 4.1 can be modified to reflect the weighting in a straightforward way. Instead of making the step size  $1/N$  for every datum, the step functions use a step size of  $W_i / \sum W_j$  for the endpoint of the  $i^{\text{th}}$  datum, where  $W_i$  is that datum's weight. The resulting p-box would be

$$[\underline{B}(X), \bar{B}(X)] = \frac{1}{\sum_{j=1}^N W_j} \left[ \sum_{i=1}^N W_i G(\underline{x}_i, X), \sum_{i=1}^N W_i G(\bar{x}_i, X) \right]$$

where  $G(Z, X)$  is one if  $X \leq Z$  and zero otherwise. This p-box bounds all of the weighted distributions associated with any possible configuration of scalars within their respective intervals in the data set.

It is also possible, in principle, to allow for interval weights, where the weights associated with the data are themselves intervals as well. However, the algorithms for this are rather complex.

## 4.2 Arithmetic mean and similar measures of central tendency

The most frequently used statistical summary is probably the mean (average). There are several different kinds of means in common use and they are all very easy to compute for interval data. This section reviews the arithmetic, geometric and harmonic means. It also discusses how weighted means can be computed.

### 4.2.1 Arithmetic mean

The derivation of a formula for the arithmetic mean of interval data  $x_i$  depends on identifying a configuration of scalar points within the respective intervals that yields the smallest possible mean, and a configuration that yields the largest possible mean. Because the mean is a monotone function, the first configuration is just the set of left endpoints  $\{\underline{x}_i\}$ , and the second configuration is just the right endpoints  $\{\bar{x}_i\}$ . The formula for the arithmetic mean of interval data  $x_i$  is therefore

$$\left[ \frac{1}{N} \sum_i \underline{x}_i, \frac{1}{N} \sum_i \bar{x}_i \right].$$

This result clearly gives us bounds on all possible configurations of values within the respective intervals, as desired.

*Numerical examples.* For the skinny data set, the arithmetic mean is the interval  $[(1.00+2.68+7.52+7.73+9.44+3.66)/6, (1.52+2.98+7.67+8.35+9.99+4.58)/6] = [5.338, 5.849]$  with outward-directed rounding. For the puffy data set, the arithmetic mean is  $[(3.5+6.9+6.1+2.8+3.5+6.5+0.15+4.5+7.1)/9, (6.4+8.8+8.4+6.7+9.7+9.9+3.8+4.9+7.9)/9] = [4.56, 7.39]$ .

The endpoints of the interval mean are the means of the bounding distributions described in section 4.1. This correspondence between the endpoints of the interval statistic and the bounding distributions does not occur for all statistics. Manski (2003) suggested the term “D-parameter” for such a statistic and the language that the statistic “respects stochastic dominance” if the bounds on the statistic are the statistics of the bounds of the p-box distribution. A statistic  $D$  is said to respect stochastic dominance if  $D(F) \leq D(G)$  for distribution functions  $F$  and  $G$  whenever  $F(X) \geq G(X)$  for all real values  $X$  (in which case the distribution  $F$  is said to “stochastically dominate”  $G$ ). Besides for means, we shall see in section 4.3 that the median and other percentiles are D-parameters and the endpoints of these statistics correspond to the respective statistic of the two bounding distributions of the p-box. But we’ll see in section 4.5 that measures of dispersion such as the variance and the interquartile range are not D-parameters and do not exhibit this correspondence. They are consequently more difficult to compute.

It seems likely that the mean of a general interval data set has been derived many times, but apparently did not appear in the literature as such until very recently. Using his approach based on a mixture model, Manski (2003, page 10) gives a formula for the mean for the case of missing data (see section 2), where some data  $x_i$  are ordinary scalars

$S_i$ ,  $i = 1 \dots N_S$ , and the rest are the vacuous interval  $x_i = [G_0, G_1]$ ,  $i = N_S + 1, \dots, N$ . In this case, bounds on the mean are

$$\frac{1}{N} \left( \sum_{i=1}^{N_S} S_i + (N - N_S)[G_0, G_1] \right) = \left[ \frac{1}{N} \sum_{i=1}^{N_S} S_i + \left( 1 - \frac{N_S}{N} \right) G_0, \frac{1}{N} \sum_{i=1}^{N_S} S_i + \left( 1 - \frac{N_S}{N} \right) G_1 \right],$$

which is a special case of the previous formula. Note that the summations are adding  $N_S$  items, but are divided by  $N$ .

You could skip the next three minor subsections addressing other measures of central tendency and resume reading with section 4.3 on page 38.

## 4.2.2 Geometric mean

The geometric mean is useful for averaging multiplicative factors such as yield coefficients. Computing the geometric mean of interval data is also straightforward. The formula is

$$\left[ \sqrt[N]{\prod_i^N \underline{x}_i}, \sqrt[N]{\prod_i^N \bar{x}_i} \right],$$

which can be applied whenever the data are nonnegative,  $0 \leq \underline{x}_i$  for all  $i$ .

*Numerical examples.* For the skinny data set, the geometric mean is the interval  $[(1.00 \cdot 2.68 \cdot 7.52 \cdot 7.73 \cdot 9.44 \cdot 3.66)^{1/6}, (1.52 \cdot 2.98 \cdot 7.67 \cdot 8.35 \cdot 9.99 \cdot 4.58)^{1/6}] = [4.186, 4.866]$ . For the puffy data set, the geometric mean is the interval  $[(3.5 \cdot 6.9 \cdot 6.1 \cdot 2.8 \cdot 3.5 \cdot 6.5 \cdot 0.15 \cdot 4.5 \cdot 7.1)^{1/9}, (6.4 \cdot 8.8 \cdot 8.4 \cdot 6.7 \cdot 9.7 \cdot 9.9 \cdot 3.8 \cdot 4.9 \cdot 7.9)^{1/9}] = [3.28, 7.09]$  with outward-directed rounding.

## 4.2.3 Harmonic mean

The harmonic mean is useful when averaging rates. Computing the harmonic mean of interval data is likewise straightforward with the formula

$$\left[ N \left( \sum_i^N \frac{1}{\underline{x}_i} \right)^{-1}, N \left( \sum_i^N \frac{1}{\bar{x}_i} \right)^{-1} \right],$$

which can also be applied whenever the data are nonnegative,  $0 \leq \underline{x}_i$  for all  $i$ .

*Numerical examples.* For the skinny data set, the harmonic mean is the  $[6/(1.00^{-1} + 2.68^{-1} + 7.52^{-1} + 7.73^{-1} + 9.44^{-1} + 3.66^{-1}), 6/(1.52^{-1} + 2.98^{-1} + 7.67^{-1} + 8.35^{-1} + 9.99^{-1} + 4.58^{-1})] = [2.978, 3.842]$ . For the puffy data set, the harmonic mean is the interval  $[9/(3.5^{-1} + 6.9^{-1} + 6.1^{-1} + 2.8^{-1} + 3.5^{-1} + 6.5^{-1} + 0.15^{-1} + 4.5^{-1} + 7.1), (9/6.4^{-1} + 8.8^{-1} + 8.4^{-1} + 6.7^{-1} + 9.7^{-1} + 9.9^{-1} + 3.8^{-1} + 4.9^{-1} + 7.9)] = [1.06, 6.74]$  with outward-directed rounding.

The harmonic mean of any data set is always less than the geometric mean of the data, which is in turn always less than the arithmetic mean of the data. We see this for the skinny data set, as  $[2.978, 3.842] < [4.186, 4.866] < [5.338, 5.849]$ , but the means for the puffy data set,  $[1.06, 6.74] \leq [3.28, 7.09] \leq [4.56, 7.39]$ , overlap each other substantially. This is a result of the greater uncertainty in the latter data set. The respective endpoints of the three intervals will always be ordered in this way (otherwise, we could tighten one or another of the intervals).

#### 4.2.4 Weighted mean

It is also straightforward to compute a weighted mean for the interval data set, where the weights  $W_i, i=1, \dots, N$ , are positive scalars, using the formula

$$M(x) \in \left[ \frac{1}{\sum_{i=1}^N W_i} \sum_i W_i \underline{x}_i, \frac{1}{\sum_{i=1}^N W_i} \sum_i W_i \bar{x}_i \right].$$

It is possible to compute a weighted mean where the weights are themselves intervals as well. The algorithms needed for this are a bit complex.

### 4.3 Median and percentiles

The median of a data set is a middle value of the collection, above and below which lie an equal number of values. Sometimes this is not a unique value. For instance, when the sample size is even, there is a range of values which could be called the median. In this situation, the average of these values is often taken as the median. When a data set includes incertitude represented by intervals, there will be uncertainty about exactly where the median is. Like the arithmetic mean, the median is a monotone function of each data value. This implies that the median respects stochastic dominance and that bounds on the median for a data set containing intervals can be evaluated by computing the medians for the left endpoints and right endpoints separately.

*Numerical examples.* The median for the skinny data set is the interval  $[\text{median}(1.00, 2.68, 3.66, 7.52, 7.73, 9.44), \text{median}(1.52, 2.98, 4.58, 7.67, 8.35, 9.99)] = [(3.66+7.52)/2, (4.58+7.67)/2] = [5.59, 6.125]$ . The median for the puffy data set is  $[\text{median}(0.15, 2.8, 3.5, 3.5, 4.5, 6.1, 6.5, 6.9, 7.1), \text{median}(3.8, 4.9, 6.4, 6.7, 7.9, 8.4, 8.8, 9.7, 9.9)] = [4.5, 7.9]$ . The medians are depicted in Figure 9 along with the p-boxes characterizing the distributions of the data sets. If we had defined median to be any value below which are 50% of the data values and above which are 50% of the data values (i.e., without any averaging), we would have computed the median for the skinny data set to be  $[[3.66, 7.52], [4.58, 7.67]] = [3.66, 7.67]$ . This is the range that we would have obtained graphically by cutting the skinny p-box at the 0.5 probability level.

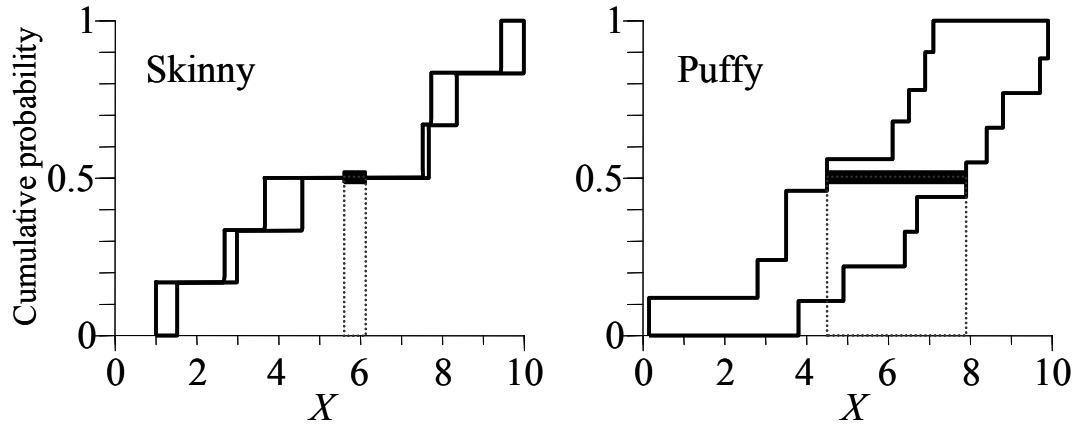


Figure 9. The medians for interval data sets.

A percentile of a data set is a number below which a given percentage of the data values lie. The median is a special case of a percentile at the 50% level. To make this definition sensible for interval data, we should say a  $P^{\text{th}}$  percentile is the interval from the *smallest* value such that  $P\%$  of the values in the data set lie below it and the *largest* value such that  $(100 - P)\%$  of the values lie above it. Like the median, percentiles interpreted this way respect stochastic dominance, so can be computed from the left and right endpoints separately. Percentiles can also be read directly from the data p-box by cutting it horizontally at the specified probability level. The algorithm to compute the  $P^{\text{th}}$  percentile is simple. First, sort the left and right endpoints of the data separately in ascending order so that  $\underline{x}_{(1)} \leq \underline{x}_{(2)} \leq \dots \leq \underline{x}_{(N)}$  and  $\bar{x}_{(1)} \leq \bar{x}_{(2)} \leq \dots \leq \bar{x}_{(N)}$ . For each datum  $i = 1, \dots, N$ , check whether  $i/N < P < (i+1)/N$ . If so, then the  $P^{\text{th}}$  percentile is  $[\underline{x}_{(i)}, \bar{x}_{(i)}]$ . If  $P = (i+1)/N$ , then the  $P^{\text{th}}$  percentile is  $[\underline{x}_{(i)}, \bar{x}_{(i+1)}]$ .

*Computability.* This algorithm to find a percentile or the median of an interval data set requires the endpoints be sorted, which requires computer time on the order of  $N \log(N)$  as the number  $N$  of samples increases. Actually, the operation requires two separate sortings which would suggest the required time is  $2 N \log(N)$ , but, by convention, the factor 2 is neglected because it isn't growing with  $N$ . The additional step of traversing the list of sorted endpoints to locate the relevant index  $i$  requires additional computer time on the order of  $N$ , which is also negligible compared to  $N \log(N)$  when  $N$  is large. Thus, this algorithm is said to require\* time on the order of  $N \log(N)$ , symbolized  $\Theta(N \log N)$ . Section 6 gives some more details about how computability is summarized.

#### 4.4 No useful estimate for mode

The mode of a data set is the most common of the data values. Unlike most other basic descriptive statistics for univariate data, the mode does not generalize to a useful summary for data sets containing interval values. The reason that the mode is

---

\*The general problem of sorting requires  $\Theta(N \log N)$  time, although there are algorithms to sort in only linear time  $\Theta(N)$  in special cases, or with special knowledge about the data. Indeed, there are also linear-time algorithms to obtain the median of a data set (Cormen et al. 2001), although such algorithms are more complicated.

problematic for interval data is really the same reason that it is problematic for real-valued data where, in principle, no two data values are likely to be the same at infinite precision. Even if intervals overlap substantially, there is always a configuration of scalar values from within the respective intervals that consist of unique values. If no values are repeated in a data set, then there is no mode, or perhaps every value is a mode. Consequently, because this is always possible, the only trustworthy bound on the mode in this case is the trivial one corresponding to the entire range of all the data.

Although we cannot compute a useful mode for an interval data set, there is another statistic we can define which may have some uses that parallel if not duplicate those of the mode. This statistic is the location on the abscissa where the difference between the left bound on the distribution function and the right bound on it is at a maximum. This is the same as the places where the most values in the data set overlap (or coincide if they're points). The formula for this statistic would be

$$\left\{ \begin{array}{l} X : \max \arg (\underline{B}(X) - \overline{B}(X)) = \max \arg \sum_{i=1}^N I(X, x_i) \\ X \in \mathbb{R} \end{array} \right\}$$

where the  $B$ 's denote the left and right bounding distributions, and the indicator function  $I(X, x_i)$  is one if  $X \in x_i$  and zero otherwise. We can call this the 'most-overlap' statistic. If there are multiple places where the difference is maximal, this statistic may not be a single interval, but could be made into one by taking its convex hull.

The most-overlap statistic is not interesting for distributions like the skinny data set where intervals do not overlap, but for intervals like the puffy data set, it shows where the overlapping among values is greatest. For the puffy data set, that location is the range  $[7.1, 7.9]$ , where five interval values overlap. This range is depicted as a black bar in Figure 10.

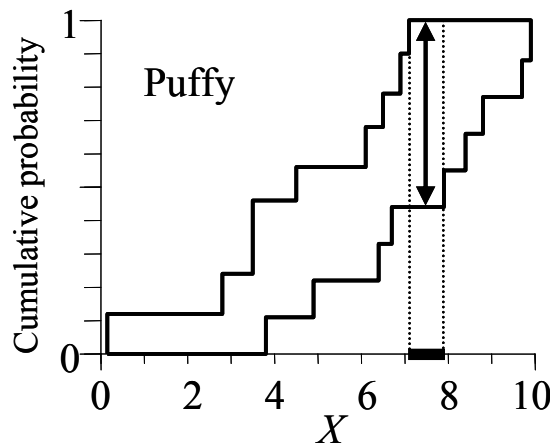


Figure 10. The most-overlap statistic.



## 4.5 Variance and other measures of dispersion

Unlike the measures of central tendency considered in section 4.2, variance and some of the other measures of dispersion considered in this section are not always easy to compute. Indeed, computing the range for the variance of interval data is provably an NP-hard problem, even if we only want to compute the variance range to within some given finite accuracy (Ferson et al. 2002a; 2002b; Kreinovich 2004). Being NP-hard means that the problem is at least as computationally difficult to solve as some problems that are widely considered by computer scientists to be generally infeasible. One basic algorithm to obtain bounds on the variance described in this section requires computational effort that grows exponentially\* with the number of intervals in the data set. Kreinovich et al. (1997) discusses the problem of NP-hardness in the context of interval uncertainty, but the central conclusion is that, for problems with a large sample size, this can make the calculation of variance prohibitive if not practically impossible. This section reviews the workable strategies for making these calculations that are available for several special cases. The algorithms described here rescue the problem from infeasibility and make practical calculations of these interval statistics in a rather wide variety of important situations.

### 4.5.1 Variance

To compute bounds on the variance for an interval data set, we must somehow identify the possible values of real numbers within each of the intervals that would in aggregate lead to the largest and smallest variances. Intuitively, it is apparent that the smallest variance arises when the points are as close as possible (given the interval ranges) to the overall mean of the data set. Conversely, the largest variance arises when the points are as dispersed as possible away from the mean. This intuition sometimes allows us to determine the extremal configurations of scalars that lead to the bounding variances by inspection. For instance, consider the skinny data set, which are replotted in Figure 11. Recall that the mean for the skinny data set was computed in section 4.2.1 to be [5.338, 5.849]. We think of a configuration of scalar point values, with each value trapped inside one of the intervals, but otherwise able to move freely. Pulling the scalars toward the mean as much as they can within their respective intervals leads to a configuration of values denoted by the diamonds in the left graph of Figure 11. Pushing the values away from the central values leads to the configuration of diamonds in the right graph. Thus, the smallest possible variance should be obtained by using the right endpoint for each interval smaller than 5 and the left endpoint for each interval larger than six. The largest variance should be obtained by selecting the opposite endpoints, that is, the left endpoints for intervals smaller than 5 and right endpoints for intervals larger than 6. These are the extremal configurations for the skinny data:

---

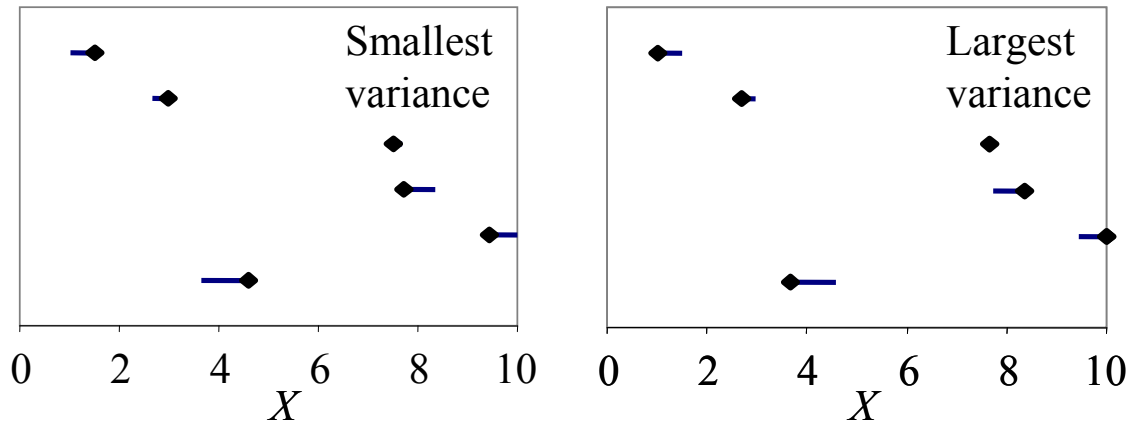
\*Mathematicians and computer scientists recognize many fine distinctions between computabilities for difficult problems, but these distinctions will not concern us here. Once a problem is NP-hard, it is already difficult enough to make it infeasible for large and possibly even moderate sample sizes. Problems whose solutions require exponential time are even worse than NP-hard.

<i>Smallest variance configuration</i>	<i>Largest variance configuration</i>
1.52	1.00
2.98	2.68
7.52	7.67
7.73	8.35
9.44	9.99
4.58	3.66

From these two configurations, one can compute bounds on the variance using either the formula for the population statistic (divide by  $N$ ) or the formula for the sample statistic (divide by  $N - 1$ ), as shown in this table:

<i>Variance</i>	<i>Population statistic</i>	<i>Sample statistic</i>
	[7.919, 10.760]	[9.503, 12.912]

These interval bounds are exact, which to say they are both rigorous in the sense that they are guaranteed to enclose the true variance assuming the data values are within their respective intervals and best possible in the sense that they are as tight as possible given those intervals. They exactly bound the variance of the data (which are imperfectly measured), which is a problem of descriptive statistics. The interval for the sample variance does not necessarily bound the variance of the underlying distribution from which those data were drawn, which is a problem of inferential statistics. However, it gives exact bounds on an estimator for that underlying variance.



**Figure 11. Configuration of scalar values (diamonds) within the skinny data set (line segments) leading to extremal variances.**

When there is a lot of overlap among the intervals, finding the extremal configurations that lead to the smallest and largest possible variances may be more difficult, although the intuition about how the extremal variances arise is the same no matter how wide the intervals are or how much overlap they exhibit. Consequently, it is still possible to build intuition about the problem by considering it from first principles. For instance, if every

interval datum overlaps a common core region of values, then the smallest possible variance for the data set is zero. Thus, we can deduce that if the intersection of all the data intervals is non-empty, which can often be determined by inspection, then the lower bound on the variance is zero.

The puffy data set has a lot of overlap among its intervals, but there is no core that every interval overlaps. In such a case, the smallest possible variance can be computed by finding a scalar value from within the interval mean—which from section 4.2.1 we know is  $[4.56, 7.39]$ —such that, when the data values are allowed to float freely within their respective intervals and come as close as possible to this special value within the mean range, their variance is minimized. Thus, in principle, we could let this special value range over many values in  $[4.56, 7.39]$ , assign values for each data interval as close as possible to this value, and then compute the variance of these assigned values. The same strategy could be used to find the largest possible variance by pushing the values within their data intervals away from the special value. Note, however, that the special value that leads to minimum variance will probably not be the same as the special value that leads to the maximum variance. The only problem is that the possible values within the mean range that we would need to try as the special values are infinitely many in number because it is a continuous range. Consequently, for large data sets with a lot of overlap among the intervals, finding the extremal configurations that lead to the smallest and largest possible variances becomes difficult.

The following subsections describe some of the important algorithms that have been recently developed for computing bounds on variance in various special cases.

#### 4.5.1.1 Lower bound on the variance

First let us consider an algorithm to compute the *lower* bound on the variance of interval data, which we symbolize as  $\underline{V}$ . We are looking for the configuration of scalar points  $\{X_i\}$  constrained to lie within their respective intervals that minimizes variance. The algorithm is based on the fact that when the variance attains a minimum on each interval  $[\underline{x}_i, \bar{x}_i]$  then there are three possibilities:

- $\partial V / \partial X_i = 0$  at the minimum somewhere in the interior of the interval, or
- the minimum is attained at the left endpoint  $X_i = \underline{x}_i$  and  $0 < \partial V / \partial X_i$ , or
- the minimum is attained at the right endpoint  $X_i = \bar{x}_i$  and  $\partial V / \partial X_i < 0$ .

Because the partial derivative is equal to  $2(X_i - M)/N$ , where  $M$  is the average of the selected  $X_i$  values, we conclude that either

- $X_i = M$ , or
- $X_i = \underline{x}_i > M$ , or
- $X_i = \bar{x}_i < M$ .

Therefore, if we knew where  $M$  is located with respect to all the interval endpoints, we could uniquely determine the corresponding minimizing value  $X_i$ , for every  $i$ .

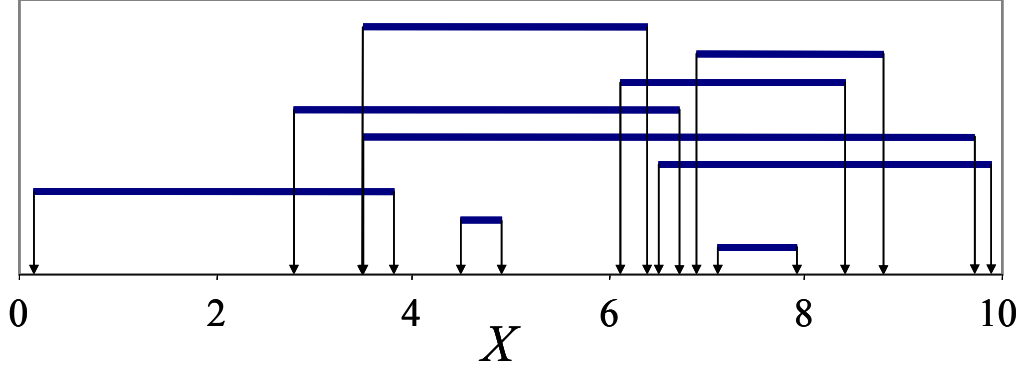
- If  $\bar{x}_i \leq M$ , then  $X_i = \bar{x}_i$ , or
- if  $M \leq \underline{x}_i$ , then  $X_i = \underline{x}_i$ ,
- otherwise  $X_i = M$ .

To find the smallest possible value of the variance, we should sort all  $2N$  endpoints  $\underline{x}_1, \bar{x}_1, \underline{x}_2, \bar{x}_2, \dots, \underline{x}_N, \bar{x}_N$  together into one sequence. We will denote the sorted list of endpoints by  $Y_1 \leq Y_2 \leq \dots \leq Y_{2N}$ , but we will still keep track of whether each is a left or right endpoint and which datum it is associated with. For each zone  $[Y_k, Y_{k+1}]$  defined by adjacent values in this sorted list, assume that  $M$  is within the zone, assign the values of  $X_i$  according to the three rules above, and compute the average and variance for these  $X_i$ . The smallest of these tentative variances for which the average actually falls within the zone is the lower bound on the true variance of the given interval data set, i.e., the variance we would observe if the data had been perfectly precise.

If the zone has no data intervals overlapping it, then the value of  $M$  is computed as the average of the  $X_i$  values assigned to be as close to the zone as possible. If the zone corresponds to the interior of one or more data intervals, then the value of  $M$  is computed as the average of the values of  $X_i$  so assigned for all intervals *other* than those overlapping the zone. And the values for those intervals that do overlap are taken to be the same as  $M$ . For example, consider the skinny data set (given on page 28 and plotted in Figure 4). The endpoints divide the  $X$ -axis into 11 zones, excluding the regions smaller than the leftmost endpoint and larger than the rightmost one. Five zones are overlapped by no data intervals. For each of these zones, we place  $M$  in it and arrange the  $X_i$  values so they are as close as possible to  $M$ . This suffices to determine  $M$  because all the  $X_i$  have been unambiguously assigned. There are also six zones that are overlapped by a data interval. For each of these six zones, we arrange the scalar value in each of the other five intervals so that it is as close as possible to the overlapped zone and compute  $M$  as the average of these five numbers.

At this point, we can check whether the value of  $M$  as now computed as the average of the  $X_i$  actually falls within the zone where it was assumed to be. If it does, then the variance associated with the configuration of scalar points  $X_i$  that produced it is a candidate for one leading to the smallest possible variance of the given interval data set. If the computed  $M$  does not fall inside that zone, then the hypothesized configuration  $X_i$  can not produce the smallest possible variance, and we should check another zone.

As another, slightly more complex example, consider the puffy data set (given on page 28 and depicted on Figure 5). Figure 12 shows how its nine data intervals specify 17 zones along the  $X$ -axis. One of these zones has zero width because adjacent values in the sorted list of endpoints happen to be identical. They are the left endpoints at 3.5 for the first and fifth intervals. For these data, all of the zones are covered by at least one of the data intervals, but several zones are overlapped by multiple data intervals. This just reduces the subset of the  $X_i$  that get averaged together to find  $M$ .



**Figure 12. Zones along the  $X$ -axis for computing the lower bound on variance for the puffy data.**

The formulas capturing this intuition are awkward looking but not really complex. If  $M$  is in zone  $[Y_k, Y_{k+1}]$ , then we implicitly know all the values of  $X_i$  and

$$M = \frac{1}{N} \left( \sum_{i: \bar{x}_i \leq Y_k} \bar{x}_i + (N - N_k)M + \sum_{j: \underline{x}_j \geq Y_{k+1}} \underline{x}_j \right)$$

where  $N_k$  is the sum of the number of right endpoints to the left of the zone ( $\bar{x}_i \leq Y_k$ ) and the number of left endpoints to the right of the zone ( $\underline{x}_j \geq Y_{k+1}$ ). Solving this equation for  $M$  yields

$$M = \frac{1}{N_k} \left( \sum_{i: \bar{x}_i \leq Y_k} \bar{x}_i + \sum_{j: \underline{x}_j \geq Y_{k+1}} \underline{x}_j \right)$$

or, if  $N_k$  is zero, which would mean that all data intervals overlap the zone so that we could assign  $X_i = M$  for all  $i$ , then we could assign  $M$  to be any any value inside the zone and immediately infer the lower bound on variance to be zero. Otherwise, once  $M$  is computed, we can compute the corresponding tentative variance  $V_k$  associated with our assumption that  $M$  is in the  $k$ th zone as  $M_k - M^2$  where  $M_k$  is the second population moment

$$M_k = \frac{1}{N} \left( \sum_{i: \bar{x}_i \leq Y_k} \bar{x}_i^2 + (N - N_k)M^2 + \sum_{j: \underline{x}_j \geq Y_{k+1}} \underline{x}_j^2 \right).$$

These expressions can be combined and simplified in the following expression for the tentative variance

$$V_k = \frac{1}{N} \left( \sum_{i: \bar{x}_i \leq Y_k} \bar{x}_i^2 + \sum_{j: \underline{x}_j \geq Y_{k+1}} \underline{x}_j^2 \right) - \frac{1}{N \cdot N_k} \left( \sum_{i: \bar{x}_i \leq Y_k} \bar{x}_i + \sum_{j: \underline{x}_j \geq Y_{k+1}} \underline{x}_j \right)^2,$$

or zero if  $N_k$  ever is. Together with the considerations about where  $M$  must lie, they justify the following algorithm.

*Algorithm for computing lower bound on variance.*

0. Data are  $x_i = [\underline{x}_i, \bar{x}_i]$ ,  $i = 1, \dots, N$ .
1. Compute the interval arithmetic mean of the data (see section 4.2.1).
2. Sort all  $2N$  endpoints  $\underline{x}_i, \bar{x}_i$  into a sequence  $Y_1 \leq Y_2 \leq \dots \leq Y_{2N}$ .
3. For each zone  $[Y_k, Y_{k+1}]$  that intersects with the interval mean, compute the values

$$N_k = \#\{i : \bar{x}_i \leq Y_k\} + \#\{j : \underline{x}_j \geq Y_{k+1}\}$$

$$S_k = \sum_{i: \bar{x}_i \leq Y_k} \bar{x}_i + \sum_{j: \underline{x}_j \geq Y_{k+1}} \underline{x}_j$$

$$M'_k = \frac{1}{N} \left( \sum_{i: \bar{x}_i \leq Y_k} \bar{x}_i^2 + \sum_{j: \underline{x}_j \geq Y_{k+1}} \underline{x}_j^2 \right)$$

where the  $\#$  function returns the cardinality of a set. (It may be convenient when computing these sums sequentially to make use of the fact that a sum for  $k > 1$  differs from the previous for  $k-1$  one by no more than two terms.)

4. If  $N_k = 0$  for any  $k$ , return  $\underline{V} = 0$  as the answer and stop the algorithm.
5. If  $N_k \neq 0$  for all  $k$ , then for each  $k$  check whether the ratio  $S_k/N_k$  is inside the  $k$ th zone  $[Y_k, Y_{k+1}]$  and also inside the interval arithmetic mean for the data set.
6. If the ratio is inside both, then compute  $V_k = M'_k - S_k^2/(N \cdot N_k)$ .
7. The smallest of these  $V_k$  is the desired bound  $\underline{V}$ .

The above algorithm is fairly fast. To obtain the lower bound on the variance of an interval data set, the algorithm needs to compute variances for at most  $2N-1$  distinct configurations of points, one for each possible zone in which  $M$  could fall. In practice, an implementation could take advantages of several shortcuts. For instance, only zones that overlap the interval estimate of the mean need to be checked. Zones of zero width can be neglected and if the computed  $M$  for any zone is outside the interval estimate of the mean for the interval data, we need not bother to finish the calculation of the tentative variance for that zone because it cannot be the smallest possible variance for those data. Also, for each zone  $k > 1$ , the tentatively calculated variance differs from the previous value by only a few terms, namely, the ones that depend on the values of  $X_i$  that shift from the right endpoint to the left endpoint as  $k$  increases. Keeping track of these can reduce the overall effort required by the calculation. Granvilliers et al. (2004) showed that the computational effort needed by this algorithm in the worst case grows with the number of samples  $N$  as a function that is proportional to  $N \log N$ . Computer scientists describe such an algorithm by saying it requires computational time on the order of  $N \log N$ , or

simply that it “runs in  $\mathcal{O}(N \log N)$  time”. This is a rather modest\* computational burden; it is the same order of growth of effort that is generally required to sort  $N$  numbers.

*Numerical example.* To reduce the chore of following the calculations, this example uses a data set with only five intervals:  $[2.1, 2.6]$ ,  $[2.0, 2.1]$ ,  $[2.2, 2.9]$ ,  $[2.5, 2.7]$ ,  $[2.4, 2.8]$ . Sorting these bounds together, we get ten points,  $Y_1 = 2.0$ ,  $Y_2 = 2.1$ ,  $Y_3 = 2.1$ ,  $Y_4 = 2.2$ ,  $Y_5 = 2.4$ ,  $Y_6 = 2.5$ ,  $Y_7 = 2.6$ ,  $Y_8 = 2.7$ ,  $Y_9 = 2.8$ , and  $Y_{10} = 2.9$ , that partition the data range into nine zones:

$$\begin{aligned} [Y_1, Y_2] &= [2.0, 2.1], & [Y_2, Y_3] &= [2.1, 2.1], & [Y_3, Y_4] &= [2.1, 2.2], \\ [Y_4, Y_5] &= [2.2, 2.4], & [Y_5, Y_6] &= [2.4, 2.5], & [Y_6, Y_7] &= [2.5, 2.6], \\ [Y_7, Y_8] &= [2.6, 2.7], & [Y_8, Y_9] &= [2.7, 2.8], & [Y_9, Y_{10}] &= [2.8, 2.9]. \end{aligned}$$

The arithmetic mean for the five data intervals is  $[2.24, 2.62]$ , so we need only keep the the four zones that have non-empty intersection with the mean. These zones are

$$\begin{aligned} [Y_4, Y_5] &= [2.2, 2.4], & [Y_5, Y_6] &= [2.4, 2.5], \\ [Y_6, Y_7] &= [2.5, 2.6], & [Y_7, Y_8] &= [2.6, 2.7], \end{aligned}$$

The first on this short list is the fourth ( $k = 4$ ) zone  $[Y_4, Y_5] = [2.2, 2.4]$ . For this zone, there is one value  $i$  for which  $\bar{x}_i \leq Y_k = Y_4 = 2.2$ . It is the value  $\bar{x}_2 = 2.1$ . There are also two values  $j$  for which  $\underline{x}_j \geq Y_{k+1} = Y_5 = 2.4$ . The first is the value  $\underline{x}_4 = 2.4$ , and the second is value  $\underline{x}_5 = 2.5$ . Thus,  $N_4 = 3$ . This leads to  $S_4 = \bar{x}_2 + \underline{x}_4 + \underline{x}_5 = 2.1 + 2.4 + 2.5 = 7.0$ , and  $M'_4 = (2.1^2 + 2.4^2 + 2.5^2) / 5 = 16.42 / 5 = 3.284$ . Proceeding to the fifth zone  $[Y_5, Y_6] = [2.4, 2.5]$ , we lose the value  $j = 5$  for which the left endpoint at 2.4 from our list of exceeded endpoints. Thus, now  $N_5 = 2$ ,  $S_5 = \bar{x}_2 + \underline{x}_4 = 2.1 + 2.4 = 4.6$ , and  $M'_4 = (2.1^2 + 2.4^2) / 5 = 10.66 / 5 = 2.132$ . When we move from the fifth zone to the sixth zone  $[Y_6, Y_7] = [2.5, 2.6]$ , we lose one more value  $\underline{x}_4 = 2.4$ , so  $N_6 = 1$ ,  $S_6 = 2.1$ , and  $M'_6 = 2.1^2 / 5 = 4.41 / 5 = 0.882$ . Finally, when we move from the sixth to the seventh zone  $[Y_7, Y_8] = [2.6, 2.7]$ , we gain a new value  $j = 1$  for which  $\underline{x}_1 = 2.6$ . Thus, here,  $N_7 = 2$ ,  $S_7 = 2.1 + 2.6 = 4.7$ , and  $M'_7 = (2.1^2 + 2.6^2) / 5 = 11.17 / 5 = 2.234$ . Computing the values of  $M$  for these four zones, we obtain  $S_4/N_4 = 7.0/3 = 2.333$  for the fourth zone,  $S_5/N_5 = 4.6/2 = 2.3$  for the fifth,  $S_6/N_6 = 2.1/1 = 2.1$  for the sixth, and  $S_7/N_7 = 4.7/2 = 2.35$  for the seventh. Of these four values only the first lies within its corresponding zone. Thus it is the only one that could contain a mean inducing the minimum variance. The tentative variance at the fourth zone is  $V_4 = M'_4 - S_4^2 / (N \cdot N_4) = 3.284 - 7^2 / (5 \cdot 3) = 0.017333$ , and this is therefore the minimum possible variance for the original five data intervals.

#### 4.5.1.2 Upper bound on the variance

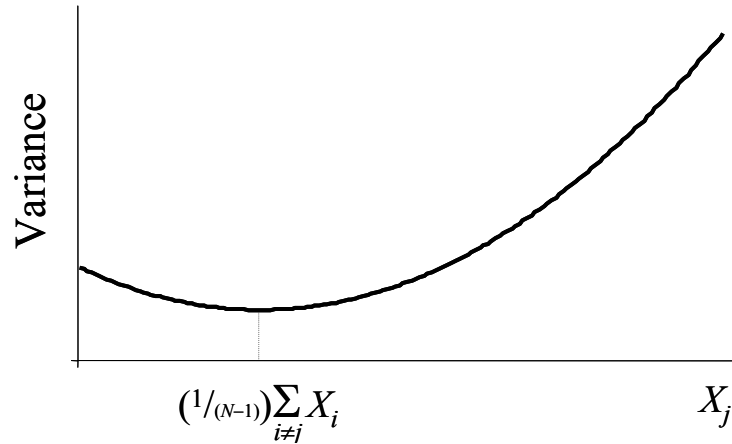
This section details several methods for computing the upper bound on variance for a data set containing intervals. A default method for the general case is reviewed first and then various more computationally convenient methods are introduced that will be applicable in different special cases.

---

\*Xiang et al. (2007) recently suggested an algorithm for computing the lower bound on variance that requires only linear time in  $N$ .

#### 4.5.1.2.1 Default method for the general case

Variance is a quadratic function of each of its data values. To see why this is so, consider the variance of a data set of scalar values  $X_i$ ,  $i=1, \dots, N$ , as a function of a particular value  $X_j$ ,  $1 \leq j \leq N$ , which might vary. The graph of this function is a parabola that takes its minimum when  $X_j$  equals the mean of the other data points, as depicted in Figure 13. As  $X_j$  deviates from this mean, the variance grows without bound. A qualitatively similar picture holds for every data value. Because variance is a quadratic function of its data and because the maximum of a quadratic function over any interval of inputs is always attained at one of the endpoints, we can, in principle, always compute the upper bound on variance  $\bar{V}$  just by computing the variance for every possible configuration of endpoints among the input intervals,  $c_1(x_1)$ ,  $c_2(x_2)$ ,  $\dots$ ,  $c_N(x_N)$ , where the  $c_i$  functions return either the left or the right endpoint of the interval. There are of course  $2^N$  of these configurations, which are the corners of the Cartesian space  $[\underline{x}_1, \bar{x}_1] \times [\underline{x}_2, \bar{x}_2] \times \dots \times [\underline{x}_N, \bar{x}_N]$ . The largest possible variance for any configuration of points inside the space occurs at one of these corners. Such a brute force search for the maximum is an algorithm that runs in  $2^N$  time, which is to say that, as the sample size  $N$  goes up, the amount of computational effort needed increases exponentially. When the number of data intervals is small, this may not be troublesome, but the necessary calculations quickly become practically impossible for problems with lots of data. As has already been mentioned, the general problem of computing the upper bound on variance for an interval data set is NP-hard (Ferson et al. 2002a; 2002b; Kreinovich 2004). This means that, in the general case, we might not be able to expect a significantly better solution than this brute force approach affords. However, there are several special cases for which much faster algorithms are available. These are reviewed in the following sections.



**Figure 13.** Variance is a quadratic function of each data value  $X_j$ .

*Numerical example.* We are, at last, equipped to compute the variance for the puffy data set. Using the algorithm described in section 4.5.1.1, we obtain the following results:

	<b>Zone</b>	$N_k$	$S_k$	$S_k / N_k$	$M'_k$	$V_k$
<b>6</b>	[ 4.5, 4.9]	5	30.4	6.08	$191.92 / 9 = 21.3244$	—
<b>7</b>	[ 4.9, 6.1]	6	35.3	5.8833	$215.93 / 9 = 23.9922$	0.91648



<b>8</b>	[ 6.1, 6.4]	5	29.2	5.84	$178.72 / 9 = 19.8578$	–
<b>9</b>	[ 6.4, 6.5]	6	35.6	5.9333	$219.68 / 9 = 24.4089$	–
<b>10</b>	[ 6.5, 6.7]	5	29.1	5.82	$177.43 / 9 = 19.7144$	–
<b>11</b>	[ 6.7, 6.9]	6	35.8	5.9667	$222.32 / 9 = 24.7022$	–
<b>12</b>	[ 6.9, 7.1]	5	28.9	5.78	$174.71 / 9 = 19.4122$	–
<b>13</b>	[ 7.1, 7.9]	4	21.8	5.45	$124.3 / 9 = 13.8111$	–

(Normal rounding is used in this table.) Eight of the seventeen zones intersected with the interval mean. But only one has the ratio  $S_k/N_k$  within the zone, so it is the only one producing a tentative variance, which is therefore the minimum variance  $\underline{V} = 0.916$ .

Using the brute force method of checking every combination of endpoints, there are  $2^9 = 512$  different configurations for which variances were combined. The largest of these was 10.975. Among the 512 corner configurations, the smallest variance observed is only 1.638. Notice that this value is *not* the smallest possible given the interval data because, unlike the maximum of a quadratic function like variance, the minimum need not occur on a corner. The configurations of scalars within the intervals of the puffy data set that produced the extreme variances are shown below and depicted in Figure 14.

<i>Smallest variance configuration</i>	<i>Largest variance configuration</i>
5.8833	3.5
6.9	8.8
6.1	8.4
5.8833	2.8
5.8833	9.7
6.5	9.9
3.8	0.15
4.9	4.5
7.1	7.9

From these bounds on population variance (defined with division by  $N$ ), one can also compute bounds on the sample variance (defined with division by  $N - 1$ ), as summarized in this table:

<i>Variance</i>	<i>Population statistic</i>	<i>Sample statistic</i>
	[0.916, 10.975]	[1.031, 12.347]

We've used outward-directed rounding to display these intervals with only three decimal places, so the last digit of both upper values is bigger than what one might otherwise expect.

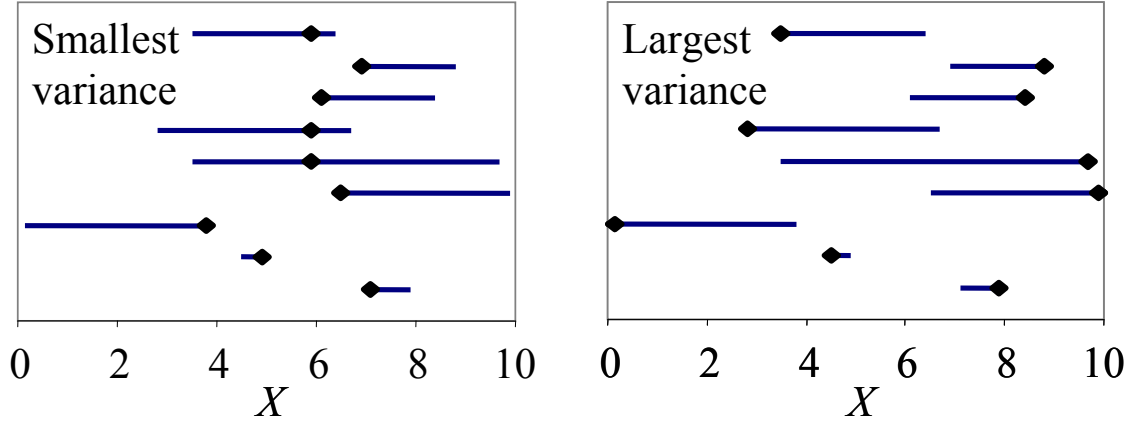


Figure 14. Configuration of scalar values within the puffy data set leading to extremal variances.

The rest of the subsections of 4.5.1 describe algorithms that are much more efficient at computing the upper bound on variance for various special cases of interval data. Unless you are particularly interested in the algorithms, you could skip this material and continue reading at section 4.5.2 on page 58.

#### 4.5.1.2.2 No-nesting data

For the case of no-nesting data, we can arrange the intervals in lexicographic order by their endpoints. In such an ordering,  $x_i \leq x_j$  if and only if either  $\underline{x}_i \leq \underline{x}_j$ , or  $\underline{x}_i = \underline{x}_j$  and  $\bar{x}_i \leq \bar{x}_j$ . Starks et al. (2004) proved that the maximum of  $V$  is always attained if, for some  $k$ , the first  $k$  values  $x_i$  are equal to  $\underline{x}_j$  and the next  $N-k$  values  $x_i$  are equal to  $\bar{x}_i$ . The proof is by reduction to a contradiction: if in the maximizing configuration  $(x_1, \dots, x_N)$ , some  $\bar{x}_i$  is smaller than some  $\underline{x}_j$ , where  $i < j$ , then we can increase  $V$  while keeping  $E$  intact, which is in contradiction of the assumption that the configuration was maximizing. Specifically, if  $\Delta_i \leq \Delta_j$ , an increase in  $V$  could be achieved by replacing  $\bar{x}_i$  with  $\underline{x}_j = \bar{x}_i - 2\Delta_i$  and  $\underline{x}_j$  with  $\underline{x}_j + 2\Delta_i$ . Otherwise, an increase could be achieved by replacing  $\underline{x}_j$  with  $\bar{x}_j = \underline{x}_j + 2\Delta_j$  and  $\bar{x}_i$  with  $\bar{x}_i - 2\Delta_j$ . As a result of these considerations, we arrive at the following algorithm.

*Algorithm for computing upper bound on variance for no-nesting interval data.*

0. Data are  $x_i = [\underline{x}_i, \bar{x}_i]$ ,  $i = 1, \dots, N$ , and  $\underline{x}_i \neq \bar{x}_i$  implies  $[\underline{x}_i, \bar{x}_i] \not\subset [\underline{x}_j, \bar{x}_j]$  for all  $i = 1, \dots, N$ , and any  $j = 1, \dots, N$ .
1. Sort the intervals  $x_i = [\underline{x}_i, \bar{x}_i]$  into lexicographic order.
2. For  $k = 0, 1, \dots, N$ , compute the value  $V = M - E^2$  for the corresponding configurations  $x^{(k)} = (\underline{x}_1, \dots, \underline{x}_k, \bar{x}_{k+1}, \dots, \bar{x}_N)$ . (Note that, when we go from a vector  $x^{(k)}$  to the vector  $x^{(k+1)}$ , only one component changes, so only one term changes in each of the sums  $E$  and  $M$ , which makes them easy and fast to compute).

3. The largest  $V$  computed in step 2 is the desired upper bound on the variance for the no-nesting interval data.

How good is this algorithm? The lexicographic ordering takes  $\mathcal{O}(N \log N)$  time.

Computing the initial values of  $E$  and  $M$  requires linear time  $\mathcal{O}(N)$ . For each  $k$ , computing the new values of  $E$  and  $M$  requires a constant number of steps, so overall, computing all  $N$  values of  $E$  and  $M$ , and hence  $V$ , requires linear time. Thus, the overall time of this algorithm is  $\mathcal{O}(N \log N)$ .

*Numerical example.* Suppose  $x_1 = [0, 3]$ ,  $x_2 = [2, 4]$ , and  $x_3 = [1, 4]$ . It is easy to check that none of these three intervals is a proper subset of the interior of any of the others, so the three intervals are a no-nesting data set. Sorting the intervals in lexicographic order and reindexing them yields  $x_1 = [0, 3]$ ,  $x_2 = [1, 4]$ , and  $x_3 = [2, 4]$ . The algorithm considers these four configurations

$$\begin{aligned} x^{(0)} &= (3, 4, 4), \\ x^{(1)} &= (0, 4, 4), \\ x^{(2)} &= (0, 1, 4), \\ x^{(3)} &= (0, 1, 2). \end{aligned}$$

For  $x^{(0)}$ , we have  $E = (3+4+4)/3 = 11/3$ , and  $M = (3^2+4^2+4^2)/3 = 41/3$ , so  $V = (41/3) - (11/3)^2 = 123/9 - 121/9 = 2/9$ . The vector  $x^{(1)}$  differs from  $x^{(0)}$  only in the first component; it has  $X_1 = \underline{x}_1 = 0$  instead of  $X_1 = \bar{x}_1 = 3$ . To get the new value of  $E$ , we can simply subtract  $(\bar{x}_1 - \underline{x}_1)/N = (3-0)/3 = 1$  from the old value, which gives  $E = 8/3$ .

Similarly, to get the new value of  $M$ , we subtract  $(3^2 - 0^2)/3 = 3$  from the old value. This gives  $M = 32/3$ . Therefore  $V = (32/3) - (8/3)^2 = 32/3 - 64/9 = 32/9$ . Next, the vector  $x^{(2)}$  differs from  $x^{(1)}$  only by the second term, so it has  $X_2 = \underline{x}_2 = 1$  instead of  $X_2 = \bar{x}_2 = 4$ .

Thus, to get the new value of  $E$ , we can simply subtract  $(\bar{x}_2 - \underline{x}_2)/N = (4-1)/3 = 1$  from the previous value, which yields  $E = 5/3$ . To get the new value of  $M$ , we can subtract  $(4^2 - 1^2)/3 = 5$  from the previous value, which gives  $M = 17/3$ . So  $V = 17/3 - (5/3)^2 = 17/3 - 25/9 = 26/9$ . Finally, the vector  $x^{(3)}$  differs from  $x^{(2)}$  only by the third term. It has  $X_3 = \underline{x}_3 = 2$  instead of  $X_3 = \bar{x}_3 = 4$ . So, to get the new value of  $E$ , we can simply subtract  $(\bar{x}_3 - \underline{x}_3)/N = (4-2)/3 = 2/3$  from the previous value, which gives  $E = 1$ . The new value of  $M$  is obtained by subtracting  $(4^2 - 2^2)/3 = 4$  from the previous value, which gives  $M = 5/3$  and  $V = 5/3 - 1^2 = 2/3$ . Of the four values of variance computed, the largest is  $32/9$ , so it is the upper bound on the variance.

#### 4.5.1.2.3 Arrangeable data

Arrangeable data sets consist of measurements that can be partitioned into  $D$  subgroups, each of which is a no-nesting data set. For arrangeable data, we can similarly prove that if we sort the intervals in each subgroup into lexicographic order, then the maximum of the variance is attained when, for intervals corresponding to each subgroup, the values  $X_i$  corresponding to this subgroup form a sequence  $(\underline{x}_1, \underline{x}_2, \dots, \underline{x}_{K_j}, \bar{x}_{K_j+1}, \bar{x}_{K_j+2}, \dots, \bar{x}_{N_j})$ , where  $N_j$  is the total number of values in the  $j$ th subgroup. To find the maximum of the variance, we must find the values  $K_1, \dots, K_D$  corresponding to the  $D$  subgroups. For

these values,  $V = M - E^2$ , where  $M = \sum M_j$  and  $E = \sum E_j$ , where  $E_j$  and  $M_j$  denote the averages of  $X_i$  and of  $X_i^2$  taken by using only data in the  $j$ th subgroup.

For each subgroup  $j$ , we can compute all  $N_{j+1}$  possible values  $E_j$  and  $M_j$  in linear time. There are fewer than  $N^D$  combinations of  $K_i$ 's. For each combination, we need  $D$  additions to compute  $E = \sum E_j$ ,  $D$  additions to compute  $M = \sum M_j$ , and a constant number of operations to compute  $V = M - E^2$ . Thus, overall, this algorithm needs  $\mathcal{O}(N^D)$  time.

#### 4.5.1.2.4 Thin data

Xiang et al. (2007a) described a *linear-time* algorithm for computing the upper bound on variance for thin interval data, i.e., interval data in which the intervals when symmetrically narrowed by dividing each width by the sample size  $N$ , none is a proper subset of the interior of another. Interval data sets that have no intersections or have the no-nesting property are special cases of thin interval data. A data set is thin if

$$|\bar{x}_i - \underline{x}_i - \bar{x}_j + \underline{x}_j| \leq N |\underline{x}_i + \bar{x}_i - \underline{x}_j - \bar{x}_j|$$

for all  $i \neq j$ . Because these ‘narrowed’ intervals have the no-nesting property, they can be ordered lexicographically without ambiguity. This also means that the midpoints of the original intervals can likewise be ordered. Starks et al. (2004; Dantsin et al. 2006) showed that, in such an ordering, the largest possible variance is attained by a configuration of scalar values within the respective intervals that occupy the left endpoints for the first  $K$  intervals and the right endpoints for the remaining intervals, for some integer  $K$ . A brute force search for the index value  $K$  that maximizes variance can be found with an algorithm that runs in  $\mathcal{O}(N \log N)$  time (Dantsin et al. 2006), but by exploiting observations about how tentative variance calculations change at nearby corners of the Cartesian space formed by the input intervals, Xiang et al. (2007a) were able to derive an even better algorithm that runs in linear time  $\mathcal{O}(N)$ , which is given below.

*Algorithm for computing upper bound on variance for thin interval data.*

0. Data are  $x_i = [\underline{x}_i, \bar{x}_i]$ ,  $i=1, \dots, N$ , and  $|\bar{x}_i - \underline{x}_i - \bar{x}_j + \underline{x}_j| \leq N |\underline{x}_i + \bar{x}_i - \underline{x}_j - \bar{x}_j|$  for all  $i \neq j$ .
1. Initialize  $I^- = \emptyset$  and  $I^+ = \emptyset$ .
2. Initialize  $I = \{i : \bar{x}_i \neq \underline{x}_i, i=1, \dots, N\}$ , the set of indices of nondegenerate intervals.
3. Set  $E^- = E^+ = 0$ .
4. Compute  $e = \sum_{i: \underline{x}_i = \bar{x}_i} x_i$ .
5. Iterate until  $I = \emptyset$  the following steps:
  - a. Compute the index  $m$  of the median of the set  $\{Z_i : Z_i = (\underline{x}_i + \bar{x}_i), i \in I\}$  and  $M = (\underline{x}_m + \bar{x}_m)$  for the corresponding interval.

- b. Partition the elements of  $I$  into two sets  $P^+ = \{i : (\underline{x}_i + \bar{x}_i) \leq M, i \in I\}$  and  $P^- = \{j : M < (\underline{x}_j + \bar{x}_j), j \in I\}$ .
- c. Compute  $e^- = E^- + \sum_{i \in P^-} \underline{x}_i$ , and  $e^+ = E^+ + \sum_{i \in P^+} \bar{x}_i$ .
- d. Compute  $e^* = e + e^- + e^+$ .
- e. Compute  $Y = (NM - \bar{x}_m + \underline{x}_m) / 2$ .
- f. If  $Y = e^*$  then replace  $I^-$  with  $I^- \cup P^-$ ,  $I^+$  with  $P^+$ , and  $I$  with  $\emptyset$ , but
- g. Otherwise
  - i. If  $Y < e^*$  then assign
    1.  $I = P^+$ ,
    2.  $I^- = I^- \cup P^-$ , and
    3.  $E^- = e^-$ , or
  - ii. Else if  $Y > e^*$  then assign
    1.  $I = P^-$ ,
    2.  $I^+ = I^+ \cup P^+$ , and
    3.  $E^+ = e^+$ .
6. Compute the variance of the vector of scalar values  $X_1, \dots, X_N$ , where  $X_i = \underline{x}_i$  for  $i \in I^-$  and  $X_j = \bar{x}_j$  for  $j \in I^+$ , and  $X_k = \underline{x}_k = \bar{x}_k$  if  $x_k$  is a degenerate interval.
7. This variance is the desired bound upper bound for the thin interval data set.

$I^-$  is the set of indices for intervals whose left endpoints we have already inferred will participate in the configuration maximizing variance, and the set  $I^+$  identifies those intervals whose right endpoints will. The set  $I$  holds the indices of the intervals about which we are still undecided. At each iteration of the algorithm, the set  $I$  is divided in half, and the values of  $E^-$  and  $E^+$  are the sums

$$E^- = \sum_{i \in I^-} \underline{x}_i, \quad E^+ = \sum_{i \in I^+} \bar{x}_i.$$

That is,  $E^-$  is the sum of all the participating left endpoints and  $E^+$  is the sum of all of the participating right endpoints. At each iteration within the algorithm, a median must be calculated, which can be done in linear time (Cormen et al. 2001), and all other operations with the elements of  $I$  require a time that is linear in its number of elements. We start with  $I$  having  $N$  elements, or less if there are degenerate intervals, and at each step it is halved. Therefore the overall computation time is no greater than  $C(N + N/2 + N/4 + \dots)$ , where  $C$  is some constant. But this sum is no larger than  $2CN$ , which means it is linear in  $N$ .

One might expect that having this  $\mathcal{O}(N)$  algorithm for thin data obviates the  $\mathcal{O}(N \log N)$  algorithm for no-nesting data described in the previous section. Not only does it require fewer computational steps, but it applies to more data sets, because all no-nesting data are special cases of thin data. The  $\mathcal{O}(N \log N)$  algorithm is not obsolete, however, because the actual *time* required for each step in these two algorithms may differ. Likewise, the  $\mathcal{O}(N \log N)$  algorithm is quite a bit simpler to implement than that for the  $\mathcal{O}(N)$  algorithm. Indeed, even checking for thinness requires a more complicated algorithm than does checking for the no-nesting property. Thus, the  $\mathcal{O}(N \log N)$  algorithm may still be useful because it requires less effort to implement and may actually require less computation time than the more sophisticated  $\mathcal{O}(N)$  algorithm, even though the former requires more steps as a function of sample size than the latter. Of course, for very large data sets, for which  $N$  is large, the  $\mathcal{O}(N)$  algorithm will eventually be preferable no matter the difference in time per step or overhead implementation costs.

#### 4.5.1.2.5 Same-precision, binned data and detect/non-detect data

The  $\mathcal{O}(N \log N)$  algorithm for no-nesting data in section 4.5.1.2.2 and the  $\mathcal{O}(N)$  algorithm for thin data in section 4.5.1.2.4 both also apply to same-precision data, binned data and detect/non-detect data, which are all special cases of no-nesting and thin data. Thus, feasible algorithms are available to compute the upper bound on variance for a wide variety of reasonably sized interval data sets that commonly arise in censuses, questionnaires, cases involving privacy restrictions, and the appearance of non-detects in chemical composition or survey studies.

#### 4.5.1.2.6 Scattered, few- and no-intersections data

Another feasible algorithm to compute the upper bound on variance is also possible for the case where no more than  $K < N$  intervals have a common intersection. As defined in section 3.5, these are few-intersections data sets, which generalize no-intersections data sets. In fact, the algorithm described below also works with scattered data, which is the still more general case. As defined in section 3.5, an interval data set is called scattered if there are no more than  $K$  intervals have a common intersection after their widths are reduced by a factor of  $1/N$ .

Ferson et al. (2005a) described an algorithm that runs in quadratic time  $\mathcal{O}(N^2)$ . It is a search for the configuration of scalar values  $X_i \in x_i$  that leads to the largest possible variance. The first step of the algorithm is to compute

$$y_i = \left[ \underline{x}_i + \bar{x}_i - \frac{(\bar{x}_i - \underline{x}_i)}{N}, \underline{x}_i + \bar{x}_i + \frac{(\bar{x}_i - \underline{x}_i)}{N} \right] / 2$$

for each  $i = 1, \dots, N$ . We can call these the ‘narrowed’ intervals. They have the same midpoints as the original data  $x_i$ , but they have much smaller widths. The endpoints of these narrowed intervals are sorted together to form a sequence  $Y_1 \leq Y_2 \leq \dots \leq Y_{2N}$ , which is then used to partition the real line into  $2N+1$  zones  $[Y_k, Y_{k+1}]$ , where  $Y_0 = -\infty$  and  $Y_{2N+1} = \infty$ . Then, for any zones that do not intersect with interval arithmetic mean of the data  $x_i$  (see section 4.2.1), we assign  $X_i = \bar{x}_i$  if  $Y_{k+1} \leq \underline{x}_i$ , or  $X_i = \underline{x}_i$  if  $\bar{y}_i \leq Y_k$ . For all other

unassigned values  $X_i$ , we consider both endpoints as possibilities. As a result, we obtain one or several configurations  $X_i$ . For each of these configurations, we check whether the average of the values is indeed within the current zone, and, if it is, compute the variance for the configuration. The largest such variance computed is returned as the desired maximal variance for the data  $x_i$ .

The proof that this algorithm only requires  $\Theta(N^2)$  time is based on the fact that for each zone, there are at most  $K$  indices  $i$  for which the  $i$ th narrowed interval  $y_i$  contains this zone and therefore, at most  $K$  indices for which we had to consider both endpoints. As a result, for each zone, there are at most  $2^K$  corresponding sequences  $X_i$  to check.

With a little work, we can derive a better algorithm that runs in only  $\Theta(N \log N)$  time. We first sort the lower endpoints of the narrowed intervals  $y_i$ . This operation requires  $\Theta(N \log N)$  time (Cormen et al. 2001). Without loss of generality, in fact, we'll assume that's been done and the lower endpoints are given in increasing order:  $y_1 \leq y_2 \leq \dots \leq y_N$ . Then, as in the previous algorithm, we sort *all* the endpoints of the narrowed intervals into a single sequence  $Y_1 \leq Y_2 \leq \dots \leq Y_{2N}$ , but we will still keep track of whether each is a left or right endpoint and which datum it is associated with. This sorting also requires  $\Theta(N \log N)$  time. The notation  $k^-(i)$  will denote the left endpoint of the  $i$ th narrowed interval and  $k^+(i)$  will denote the right endpoint.

The next step of the algorithm is to produce, for each of the resulting zones  $[Y_k, Y_{k+1}]$ , the set  $S_k$  of all the indices  $i$  for which the  $i$ th narrowed interval  $y_i$  contains the zone. As already mentioned, for each  $i$ , we know the value  $k = k^-(i)$  for which  $y_i = Y_k$ . So, for each  $i$ , we place  $i$  into the set  $S_{k-(i)}$  corresponding to the zone  $[Y_{k-(i)}, Y_{k-(i)+1}]$ , into the set corresponding to the next zone, etc., until we reach the zone for which the upper endpoint is exactly  $y_i$ .

Here, we need one computational step for each new entry of  $i$  into the set corresponding to a new zone. Therefore, filling in all these sets requires as many steps as there are items in all these sets. For each of  $2N+1$  zones, as we have mentioned, there are no more than  $K$  items in the corresponding set; therefore, overall, all the sets contain no more than  $K \cdot (2N+1) = \Theta(N)$  steps. Thus, this calculation requires  $\Theta(N)$  time.

The fourth step of the algorithm is to compute, for all integers  $P$  from 0 to  $N$ , the sums

$$E_P = \frac{1}{N} \sum_{i=1}^P x_i + \frac{1}{N} \sum_{i=P+1}^N \bar{x}_i,$$

$$M_P = \frac{1}{N} \sum_{i=1}^P x_i^2 + \frac{1}{N} \sum_{i=P+1}^N \bar{x}_i^2.$$

The values are computed sequentially. Once we know  $E_P$  and  $M_P$ , we can compute  $E_{P+1} = E_P + \frac{x_{P+1} - \bar{x}_{P+1}}{N}$ , and  $M_{P+1} = M_P + \frac{(x_{P+1} - \bar{x}_{P+1})^2}{N}$ . The transition from  $E_P$  and

$M_P$  to  $E_{P+1}$  and  $M_{P+1}$  requires a constant number of computational steps, so, overall, we need  $\Theta(N)$  steps to compute all the values  $E_P$  and  $M_P$ .

The fifth step of the algorithm is to compute, for each zone  $k$ , the corresponding values of the variance. For that, we first find the smallest index  $i$  for which  $Y_{k+1} \leq y_i$ . We will denote this value  $i$  by  $p(k)$ . Since the values  $y_i$  are sorted, we can find it by using bisection (Cormen 2001). It is known that bisection requires  $\Theta(\log N)$  steps, so finding such  $p(k)$  for all  $2N+1$  zones requires  $\Theta(N \log N)$  steps.

Once  $i \geq p(k)$ , then  $y_i \geq y_{p(k)} \geq Y_{k+1}$ . So, in accordance with the justification for the quadratic-time algorithm, we should assign  $X_i = \bar{x}_i$ , as in the sums  $E_{p(k)}$  and  $M_{p(k)}$ . In accordance with the same justification, the only values  $i < p(k)$  for which we may also assign  $X_i = \bar{x}_i$  are the values for which the  $i$ th narrowed intervals contains this zone.

These values are listed in the set  $S_k$  of no more than  $K$  such intervals. So, to find all possible values of  $V$ , we can do the following. We consider all subsets  $s \subseteq S_k$  of the set  $S_k$ ; there are no more than  $2^K$  such subsets. For each subset  $s$ , we replace, in  $E_{p(k)}$  and  $M_{p(k)}$ , values  $x_i$  and  $(x_i)^2$  corresponding to all  $i \in s$ , with, correspondingly,  $\bar{x}_i$  and  $(\bar{x}_i)^2$ . Each replacement means subtracting no more than  $K$  terms and then adding no more than  $K$  terms, so each computation requires no more than  $2K$  steps. Once we have  $E$  and  $V$  corresponding to the subset  $s$ , we can check whether  $E$  belongs to the analyzed zone and, if so, compute  $V = M - E^2$ . For each subset, we need no more than  $2K+2$  computations, so for all no more than  $2^K$  subsets, we need no more than  $(2K+2) \cdot 2^K$  computations. For a fixed  $K$ , this value does not depend on  $N$ , so this means we need only a constant amount of calculation  $\Theta(1)$  steps for each zone. To perform this computation for all  $2N+1$  zones, we need  $(2N+1) \cdot \Theta(1) = \Theta(N)$  time.

The last step of the algorithm is to find the largest of the resulting values  $V$ , which will be the desired value upper bound on the variance of the data  $x_i$ . Finding the largest of  $\Theta(N)$  values requires  $\Theta(N)$  steps. Summing up all the computational effort required by this algorithm gives  $\Theta(N \log N) + \Theta(N \log N) + \Theta(N) + \Theta(N) + \Theta(N \log N) + \Theta(N) = \Theta(N \log N)$  steps. This proves the computes the upper bound on variance in  $\Theta(N \log N)$  time.

#### 4.5.1.3 When only some intervals are nondegenerate

In some situations, most of the data collected are considered precise and are represented as scalar values. One example of this is when we have a mostly complete data set that is contaminated with only a few missing values. This situation can be exploited to simplify the calculations. Suppose that, in a data set with  $N$  values, only  $H \ll N$  are nondegenerate intervals. This section briefly outlines the economies that can be enjoyed in the methods described so far for computing variance.

Finding the lower bound on variance as outlined in section 4.5.1.1 would only require determining  $X_i$  for the  $H$  nondegenerate intervals, which would mean only having to



check  $2H$  zones formed by their endpoints. The calculations would require  $\mathcal{O}(H \log H)$  time to sort the  $H$  endpoints, and  $\mathcal{O}(H)$  time to compute most of the sequential values  $S_k$ ,  $N_k$  and  $M'_k$  and their implied  $V_k$ , although they would still use  $\mathcal{O}(N)$  time to produce the initial values of  $S_k$ ,  $N_k$  and  $M'_k$ . Thus, the overall time for this exercise would be only  $\mathcal{O}(H \log H + N)$ , which will typically be much smaller than  $\mathcal{O}(N \log N)$ .

Computing the upper bound on the variance would also benefit substantially by having to deal with only  $H$  intervals. In the general case, which was described in section 4.5.1.2.1, the overall computational time would only be  $\mathcal{O}(N + 2^H)$  rather than the much larger  $\mathcal{O}(2^N)$ . For few-intersections an scattered data (discussed in section 4.5.1.2.6), the calculations would require only  $\mathcal{O}(H \log H + N)$  time. For the case of arrangeable data (discussed in section 4.5.1.2.3) arising from  $D$  measuring instruments, the calculations would require  $\mathcal{O}(H \log H)$  steps for sorting,  $\mathcal{O}(N)$  steps for the original arrangement, and  $H^D$  steps to compute the values for all  $H^D$  possible combinations of  $K_i$ . If  $2 \leq D$ , then  $H \log H \leq H^D$ , so the overall computational time would be  $\mathcal{O}(N + H^D)$ .

#### 4.5.1.4 Cheap conservative bounds on variance

One reason that few investigators have studied the implications of interval uncertainty for routine statistics may be that it is, in general, extremely hard (Ferson et al. 2002a; 2002b; Kreinovich 2004) to compute the variance or related measures of dispersion and that no approximate methods are available. This section offers an approximate method for quantitatively computing variance for interval data sets. The result is rigorous in the sense that it is sure to enclose the true variance, but the upper bound may be somewhat larger than it needs to be.

Recall that the lower bound on variance can always be computed using the algorithm described in section 4.5.1.1 which should be workable for almost any data set since it runs in  $\mathcal{O}(N \log N)$  time (the same effort required to sort a list of  $N$  values). It is the *upper* bound on variance that can be difficult. In this approach it is computed separately using an idea suggested by Roger Nelsen.

Suppose (without loss of generality) that we have separately sorted the left endpoints of all the intervals  $\{\underline{x}_1, \underline{x}_2, \dots, \underline{x}_N\}$  and the right endpoints  $\{\bar{x}_1, \bar{x}_2, \dots, \bar{x}_N\}$  of all the intervals so that  $\underline{x}_i \leq \underline{x}_j$  and  $\bar{x}_i \leq \bar{x}_j$  whenever  $1 \leq i < j \leq N$ . Compute the variance for the vector  $\{\underline{x}_1, \underline{x}_2, \dots, \underline{x}_K, \bar{x}_{K+1}, \bar{x}_{K+2}, \dots, \bar{x}_N\}$  for all integers  $K$  between 1 and  $N$ . The largest variance of these  $N-2$  computations is a conservative estimate of the upper bound of the variance of the interval data. This algorithm is effectively tracing out a hypothetical distribution that matches with the left side of the data's empirical distribution (section 4.1) for low probabilities and then matches with the right side of the empirical distribution for high probabilities. The discontinuous jump from the left side to the right side occurs at the spot that maximizes the resulting variance. This distribution might be called the dispersive distribution because it has the largest dispersion that could exist for any distribution inside the p-box.

*Numerical examples.* Let us compute the conservative upper bounds on variance for the skinny and puffy data sets. We saw in section 4.1 that the left side of the empirical distribution for the skinny data set is the distribution  $\{1.00, 2.68, 3.66, 7.52, 7.73, 9.44\}$  and its right side is  $\{1.52, 2.98, 4.58, 7.67, 8.35, 9.99\}$ . By checking the variance of the distribution formed by taking the  $K$  first elements of the left side and the  $N-K$  last elements from the right side, we find that the  $K = 3$  yields the distribution, i.e.,  $\{1.00, 2.68, 3.66, 7.67, 8.35, 9.99\}$ , with the largest population variance, which is 10.760 (with upward-directed rounding). Likewise, the p-box for the puffy data set is the pair of distributions  $\{0.15, 2.8, 3.5, 3.5, 4.5, 6.1, 6.5, 6.9, 7.1\}$  and  $\{3.8, 4.9, 6.4, 6.7, 7.9, 8.4, 8.8, 9.7, 9.9\}$ . For this data set,  $K = 4$  leads to  $\{0.15, 2.8, 3.5, 3.5, 7.9, 8.4, 8.8, 9.7, 9.9\}$  whose population variance 11.448 (with upward rounding) is the largest possible for any value of  $K$ . The dispersive distributions are depicted in Figure 15. The upper bounds on variance computed this way are conservative (i.e., no smaller than) the exact upper bounds computed previously. In the case of the skinny data, the results are just the same. In the case of the puffy data, the conservative bound is about 0.5 larger than it absolutely needs to be.

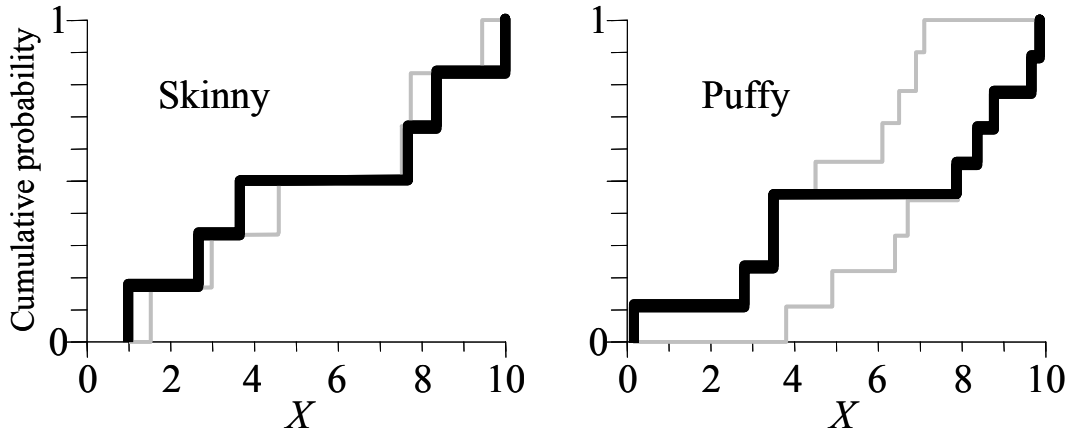


Figure 15. Empirical distributions (gray) and dispersive distributions (black) that lead to conservative upper bounds on variance.

#### 4.5.2 Standard deviation and standard error

The standard deviation  $s$  of an interval data set can be computed immediately from the variance  $v$  (section 4.5.1) simply by taking its square root,

$$s = [\underline{s}, \bar{s}] = \sqrt{v} = [\sqrt{\underline{v}}, \sqrt{\bar{v}}].$$

When the variance is a nondegenerate interval, the standard deviation will be too.

*Numerical examples.* The sample standard deviation for the skinny data set is  $[\sqrt{9.503}, \sqrt{12.912}] = [3.082, 3.594]$ . (Remember that intervals are outward rounded as described in section 3.2, so that, although the sample standard deviation is computed to be  $[3.0827\dots, 3.5933\dots]$ , the left endpoint rounds *down* and the right endpoint rounds *up* to the interval reported here.) The population standard deviation for the skinny data set is  $[\sqrt{7.919},$

$\sqrt{10.760} = [2.814, 3.281]$ . For the puffy data set, the sample standard deviation is  $[\sqrt{1.031}, \sqrt{12.347}] = [1.015, 3.514]$ , and the population standard deviation is  $[\sqrt{0.916}, \sqrt{10.975}] = [0.957, 3.313]$ .

The standard error, which is the standard deviation associated with the mean, can also be computed from the sample variance immediately by scaling for the sample size,

$$s_X = [\underline{s}_X, \bar{s}_X] = [\underline{s} / \sqrt{N}, \bar{s} / \sqrt{N}] = \sqrt{v / N} = [\sqrt{\underline{v} / N}, \sqrt{\bar{v} / N}].$$

Again, whenever the variance is a nondegenerate interval, the standard error will be too.

*Numerical examples.* The standard error for the skinny data set, i.e., the standard deviation of its mean, is  $[\sqrt{9.503}, \sqrt{12.912}] / \sqrt{6} = [1.258, 1.467]$ . For the puffy data set, the standard error is  $[\sqrt{1.031}, \sqrt{12.347}] / \sqrt{9} = [0.338, 1.172]$ .

### 4.5.3 Range and interquartile range

The range of a data set that contains intervals is the tightest interval in which all possible data values lie. It can be evaluated simply as the interval between the smallest left bound and largest right bound in the data set,

$$\left[ \min_{1 \leq i \leq N} \underline{x}_i, \max_{1 \leq i \leq N} \bar{x}_i \right].$$

The interquartile range for a data set containing intervals is the interval from the smallest value such that 25% of the values in the data set lie below it and the largest value such that 75% of the values lie above it. It is very easy to evaluate the interquartile range, and it can be done graphically. The interquartile range is often used as a characterization of dispersion because it is less sensitive to outliers than the standard deviation (or range), just as the median is less sensitive than the mean. Such characterizations are called robust statistics (Huber 1981).

Both the range and the interquartile range are sometimes expressed as *displacements* in the  $X$  direction, rather than as intervals of possible  $X$ -values between percentiles. These statistics say how wide the scatter of data is, rather than where it is. The formula for the displacement range for an interval data set is

$$\left[ \max \left( 0, \max_{1 \leq i \leq N} \underline{x}_i - \min_{1 \leq i \leq N} \bar{x}_i \right), \max_{1 \leq i \leq N} \bar{x}_i - \min_{1 \leq i \leq N} \underline{x}_i \right] = p_{100} - p_0,$$

where  $p_{100}$  and  $p_0$  are the interval estimates of the 100<sup>th</sup> and 0<sup>th</sup> percentiles respectively (see section 4.3) and the minus sign between them denotes the interval difference (section 3.1). When there are intervals in the data, even the displacement range will still generally be an interval, because we cannot be sure what the span is between the smallest

datum and the largest datum. The displacement version of the interquartile range can be evaluated as

$$p_{75} - p_{25} ,$$

where  $p_{75}$  and  $p_{25}$  are the interval estimates of the 75<sup>th</sup> and 25<sup>th</sup> percentiles respectively, and any negative values are truncated to zero. These displacement statistics do not respect stochastic dominance, so evaluating them requires examining the intervals together, rather than their sets of endpoints separately.

*Numerical examples.* The range for the skinny data set is [1, 9.99], and the range for the puffy data set is [0.15, 9.9]. The interquartile range for the skinny data set is [2.68, 8.35], and the interquartile range for the puffy data set is [3.5, 8.8], as depicted in Figure 16 along with the p-boxes characterizing the distributions of the data sets. The displacement range for the skinny data is the interval [7.92, 8.99]. The largest possible displacement range of 8.99 occurs when the realized value in the data interval [1.00, 1.52] is 1 and the realized value within the data interval [9.44, 9.99] is 9.99. This is the difference in the  $X$ -direction between the upper, far-right corner of the p-box and the lower, far-left corner. The smallest possible displacement range of 7.92 occurs when these data intervals take on the realized values 1.52 and 9.44 respectively. This is the difference between the upper, left corner and the lower, right corner of the p-box. The displacement interquartile range for the skinny data is [4.75, 5.67]. The displacement range for the puffy data is [3.3, 9.75] and its displacement interquartile range is [0.1, 5.3].

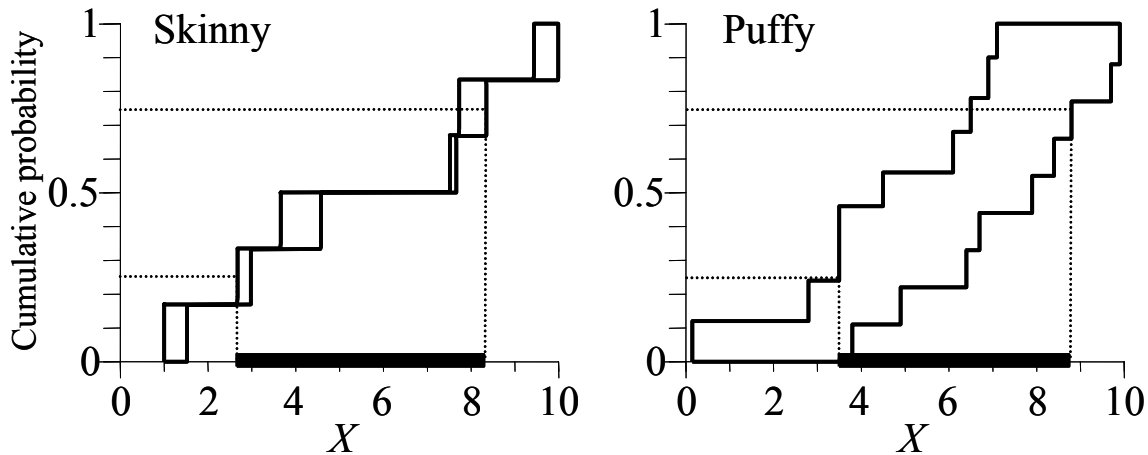


Figure 16. Interquartile ranges (as set of possible values) for interval data sets.

You may want to skip the next two sections which discuss methods to compute statistical moments and skewness. Resume reading about confidence intervals in section 4.8 on page 66.

## 4.6 Moments

This section outlines the calculation of moments about the origin and central moments about the mean for data sets that may contain intervals.

#### 4.6.1 Moments about the origin

The  $Q^{\text{th}}$  moment about the origin of a data set is the average of the  $Q^{\text{th}}$  powers of the individual data, where  $Q$  is a nonnegative integer. When  $Q$  is odd, the moment function is monotone and it respects stochastic dominance, so the  $Q^{\text{th}}$  moment is

$$\left[ \frac{1}{N} \sum_i \underline{x}_i^Q, \frac{1}{N} \sum_i \bar{x}_i^Q \right], \text{ if } Q \text{ is odd.}$$

The zeroth moment is identically one, even for interval data sets. The first moment is the arithmetic mean of the data set (section 4.2.1).

When  $Q$  is even, the moment function is not monotone, unless the data happen to all have constant sign. In the general case, the formula becomes

$$\frac{1}{N} \left( \sum_i [\underline{x}_i, \bar{x}_i]^Q \right), \text{ if } Q \text{ is even}$$

where

$$[\underline{x}, \bar{x}]^Q = \begin{cases} [\bar{x}^Q, \underline{x}^Q] & \text{if } \bar{x} \leq 0 \\ [0, \max(\underline{x}^Q, \bar{x}^Q)] & \text{if } \underline{x} \leq 0 \leq \bar{x} \\ [\underline{x}^Q, \bar{x}^Q] & \text{if } 0 \leq \underline{x} \end{cases}$$

is the integral-power function for intervals, which accounts for intervals that are negative, positive and straddle zero.

*Numerical examples.* Because both data sets consist of strictly positive values, all the moments for the skinny and puffy data sets can be computed easily and exactly. Some results are shown in the following table.

	<i>Skinny</i>	<i>Puffy</i>
0 <sup>th</sup> moment	1	1
1 <sup>st</sup> moment	[ 5.338, 5.849 ]	[ 4.56, 7.39 ]
2 <sup>nd</sup> moment	[ 37.832, 43.420 ]	[ 25.56, 58.54 ]
3 <sup>rd</sup> moment	[ 299.609, 359.409 ]	[ 154.09, 487.30 ]
4 <sup>th</sup> moment	[ 2490.269, 3134.390 ]	[ 972.13, 4200.83 ]

#### 4.6.2 Central moments of even order

The  $Q^{\text{th}}$  central moment, or central moment of order  $Q$  is defined as

$$M_Q = \frac{1}{N} \sum_{i=1}^N \left( x_i - \frac{1}{N} \sum_{j=1}^N x_j \right)^Q,$$

which is to say, it is the average  $Q^{\text{th}}$  power of the deviations from the mean. The moment is of even order if  $Q$  is even, and otherwise it is odd.

To evaluate the  $M_Q$  associated with an interval data set, we will first consider the lower bound. The partial derivative

$$\frac{\partial M_Q}{\partial x_i} = \frac{Q}{N} (x_i - E)^{Q-1} - \frac{QM_{Q-1}}{N},$$

where  $E$  denotes the true scalar arithmetic mean of the data (which may be unknown), is nonnegative if and only if  $\mu \leq x_i$  for all  $i$ , where

$$\mu \equiv E + (QM_{Q-1})^{1/(Q-1)}.$$

Thus, once we know where  $\mu$  is located with respect to the endpoints, we can find the configuration of the scalars  $X_i \in x_i$  that minimize  $M_Q$ , and compute  $\mu$  itself from its definition above evaluated at that configuration. The value  $M_{Q-1}$  can be computed based on the corresponding population moments up to  $Q-2$ . Once we know the moments for one zone, computing the moment for the next zone requires a constant-time calculation. Therefore, to find the lower bound  $\underline{M}_Q$  involves sorting the endpoints, computing the original values of the moments, and computing the moments for all zones to identify the smallest one.

Evaluating the upper bound  $\overline{M}_Q$  is NP-hard in general, but feasible algorithms are possible for a variety of the special kinds of interval data set described in section 3.5. For instance, for no-intersection or few-intersections data sets, a similar analysis of partial derivatives leads straightforwardly to an  $\mathcal{O}(N \log N)$  algorithm for the upper bound.

For a same-precision data set in which all the nondegenerate intervals have the same width, the problem of finding the upper bound is like that of the variance. In this case, the maximum is always attained at a corner of the data set  $(\underline{x}_{(1)}, \dots, \underline{x}_{(k)}, \bar{x}_{(k+1)}, \dots, \bar{x}_{(N)})$  for some ordering of the indices. The following arguments work not only for central moments, but also for generalized central moments

$$M_\psi = \frac{1}{N} \sum_{i=1}^N \psi(x_i - E)$$

for an arbitrary convex function  $\psi(X) \geq 0$  for which  $\psi(0) = 0$  and  $\psi''(X) > 0$  for all  $X \neq 0$ . Let us first show that the maximum cannot be attained in an interior of any of the intervals. In this case, at the maximizing point the first derivative

$$\frac{\partial M_\psi}{\partial x_i} = \frac{1}{N} \psi'(x_i - E) - \frac{1}{N^2} \sum_{j=1}^N \psi'(x_j - E)$$

should be equal to zero and the second derivative

$$\frac{\partial^2 M_\psi}{\partial x_i^2} = \frac{1}{N} \psi''(x_i - E) \left(1 - \frac{2}{N}\right) + \frac{1}{N^3} \sum_{j=1}^N \psi''(x_j - E)$$

is non-positive. Because the function  $\psi$  is convex,  $\psi''(X) \geq 0$ , so this second derivative is a sum of nonnegative terms, and the only case when it is nonnegative is when all these terms are zeros, that is, when  $x_j = E$  for all  $j$ . In this case,  $M_\psi = 0$  which, for nondegenerate intervals, is clearly not the largest possible value of  $M_\psi$ . Therefore, for every  $i$ , the maximum of the generalized central moment is attained either when  $x_i = \underline{x}_i$  or when  $x_i = \bar{x}_i$ . This proves the maximum is at a corner of the data set, and thus that the general case of searching for the maximum, at worst, requires  $\mathcal{O}(2^N)$  time.

We will now prove that the maximum is always attained for one of the vectors  $(\underline{x}_{(1)}, \dots, \underline{x}_{(k)}, \bar{x}_{(k+1)}, \dots, \bar{x}_{(N)})$ . To prove this, we need to show that if  $x_i = \bar{x}_i$  and  $x_j = \underline{x}_j$  for some  $i < j$  (and  $\underline{x}_i \leq \underline{x}_j$ ), then the change described in that proof, while keeping the average  $E$  intact, increases the value of  $M_\psi$ . Without loss of generality, we can consider the case  $\Delta_i \leq \Delta_j$ . In this case, the fact that  $M_\psi$  increase after the above-described change is equivalent to  $\psi(\underline{x}_j + 2\Delta_i - E) + \psi(\underline{x}_i - E) \leq \psi(\underline{x}_i - E) + \psi(\underline{x}_j + 2\Delta_j - E)$ , i.e., that  $\psi(\underline{x}_i + 2\Delta_i - E) - \psi(\underline{x}_i - E) \leq \psi(\underline{x}_j + 2\Delta_j - E) - \psi(\underline{x}_j - E)$ . Since  $\underline{x}_i \leq \underline{x}_j$  and  $\underline{x}_i - E \leq \underline{x}_j - E$ , this can be proven if we show that for every  $\Delta > 0$  (and, in particular, for  $\Delta = 2\Delta_i$ ), the function  $\psi(x + \Delta) - \psi(x)$  is increasing. Indeed, the derivative of this function is equal to  $\psi'(x + \Delta) - \psi'(x)$ , and because  $\psi''(x) \geq 0$ , we do have  $\psi'(x + \Delta) \geq \psi'(x)$ . The idea behind the algorithm is that finding the upper bound of  $M_\psi$  only requires checking all  $N$  vectors of the type  $(\underline{x}_{(1)}, \dots, \underline{x}_{(k)}, \bar{x}_{(k+1)}, \dots, \bar{x}_{(N)})$ .

*Computability.* For finding the lower bound of the central moment, sorting the endpoints requires  $\mathcal{O}(N \log N)$  time, computing the original values of the moments takes  $\mathcal{O}(N)$  time, and computing the moments for all zones takes  $\mathcal{O}(N)$  time. So the overall effort to compute the lower bound requires  $\mathcal{O}(N \log N)$  time. For computing the upper bound, computability varies for the different data cases. The algorithm for both no-intersections data sets and few-intersections data sets requires  $\mathcal{O}(N \log N)$  time. For the case of same-precision data, checking all  $N$  vectors requires  $\mathcal{O}(N \log N)$  steps. An algorithm for a several-instruments data set originating from  $M$  measuring instruments (producing intervals of  $M$  different widths) would require  $\mathcal{O}(N^M)$  time. When only  $H$  out of  $N$  intervals are nondegenerate, there can also be some economy. In this case, the algorithms

for finding the lower and upper bounds on a central moment for the general case would need, respectively, only  $\Theta(N+H \log H)$  time instead of  $\Theta(N \log N)$  time, and only  $\Theta(N+2^H)$  time instead of  $\Theta(2^N)$  time.

### 4.6.3 Central moments of odd order

As explained in the previous section, the  $Q^{\text{th}}$  central moment is a central moment of odd order if  $Q$  is odd. The first central moment is always zero, even if the inputs are intervals. The derivative has the same form  $\partial M_Q / \partial X_i = (Q/N) (X_i - E)^{Q-1} - (Q/N) M_{Q-1}$ , but due to the fact  $Q$  is odd, it leads to a more complex description of the condition  $\partial M_Q / \partial X_i \geq 0$ . This condition is equivalent to  $X_i \geq \mu^+$  or  $X_i \leq \mu^-$ , where

$$\mu^\pm \equiv E \pm (Q M_{Q-1})^{1/(Q-1)}.$$

Thus, to find all the values  $X_i$  in the configuration that extremizes the statistic, instead of knowing a single zone where  $\mu$  lies, we now need to know *two* zones: a zone containing  $\mu^-$  and a zone containing  $\mu^+$ . There are  $\Theta(N^2)$  such pairs of zones, and each needs to be tried.

*Computability.* For general interval data sets, computing both  $\underline{M}_Q$  and  $\overline{M}_Q$  with this approach would require  $\Theta(2^N)$  time. If  $K$  or fewer intervals intersect, only  $\Theta(N^2)$  time is needed. If only  $H$  out of  $N$  intervals are nondegenerate, then we need  $\Theta(N+2^H)$  time instead of  $\Theta(2^N)$ , and  $\Theta(N+H^2)$  instead of  $\Theta(N^2)$ .

*Numerical examples.* The third central moment for the skinny data is  $[-6.963, 2.556]$ . The smallest possible value of the third central moment arises from the configuration vector (1, 2.98, 7.67, 8.35, 9.44, 4.58); the largest arises from the configuration vector (1.52, 2.68, 7.52, 7.73, 9.99, 3.66). The third central moment for the puffy data set is  $[-30.435, 7.871]$ . The associated extremal configurations are (6.4, 8.8, 8.4, 6.7, 9.7, 9.823871978, 0.15, 4.9, 7.9) and (3.5, 6.9, 6.1, 2.8, 3.5, 9.9, 3.8, 4.5, 7.1). Notice that all but one of the values in these extremal configurations are endpoints from the original interval data. The one that isn't has many decimal places. The bounds and extremal configurations were found using Microsoft Excel's Solver facility (Fylstra et al. 1998) which uses simplex, generalized-reduced-gradient, and branch-and-bound methods to solve optimization problems.

## 4.7 Skewness

Skewness is a measure of the asymmetry of a distribution. A distribution is said to be positively skewed, or right-skewed, if its right tail is longer than symmetry would suggest, and negatively or left-skewed if its left tail is longer. The standardized moment of skewness is

$$\gamma_1 = \mu_3 / \sigma^3$$



where the numerator  $\mu_3$  denotes the third central moment about the mean and denominator  $\sigma^3$  denotes the cube of the standard deviation. Karl Pearson also suggested an alternative measure of skewness as

$$\frac{\text{mean} - \text{median}}{\text{standard deviation}}$$

although this measure does not generally agree with the standardized moment of skewness.

As we saw with the variance in section 4.5.1, the optimal calculation of the skewness coefficient for data sets containing intervals will be considerably more complicated than it would be for a data set with only point estimates. As yet, little attention has been devoted to algorithms that could be used to compute exact bounds on either of these coefficients. However, it is possible to use naive interval analysis to obtain an enclosure of the desired interval. This enclosure is guaranteed to bound the true skewness of the data set, but probably will not be optimally tight. Alternatively, an inside-out approximation can be obtained by selecting scalar values from each of the intervals and computing the skewness coefficient. As this process is repeated for different possible configurations of point values within the respective intervals, the smallest and largest skewnesses observed characterize a range that must lie inside the true interval for skewness given the uncertainty expressed in the data intervals. The combination of the enclosure obtained by naive interval analysis and what might be called the ‘inner estimate’ obtained by sampling may be informative.

*Numerical examples.* As computed in section 4.6.3, the third central moments for the skinny and puffy data sets are, respectively, the intervals  $[-6.963, 2.556]$  and  $[-30.435, 7.871]$ . We saw in section 4.5.2 the corresponding population standard deviations to be the intervals  $[2.814, 3.281]$  and  $[0.957, 3.313]$ . From these results we can compute a sure (but probably not optimal) bound on the standardized moment of skewness for the skinny data set as  $[-6.963, 2.556] / [2.814, 3.281]^3 = [-0.313, 0.115]$ . The same statistic for the puffy data set is  $[-30.435, 7.871] / [0.957, 3.313]^3 = [-34.725, 8.981]$ . As computed in sections 4.2.1 and 4.3, the arithmetic means for the skinny and puffy data sets are, respectively, the intervals  $[5.338, 5.849]$  and  $[4.56, 7.39]$ , and the medians are  $[5.59, 6.125]$  and  $[4.5, 7.9]$ . Now we can compute bounds on Pearson’s measure of skewness for the skinny data as  $([5.338, 5.849] - [5.59, 6.125]) / [2.814, 3.281] = [-0.279, 0.093]$ , and for the puffy data as  $([4.56, 7.39] - [4.5, 7.9]) / [0.9573, 3.3128] = [-3.489, 3.019]$ . Because of the repeated parameters involved in these formulations, the resulting intervals may be wider than is necessary to contain all the skewness values that might arise from the interval data. Monte Carlo simulations involving 100,000 sample vectors selected at random from the interval data sets were used to find inner bounds for the standardized and Pearson skewnesses. The results of the simulation are given in the table along with the intervals just computed using naive interval calculations:

<i>Skinny</i>	Standardized skewness	Pearson's skewness
Monte Carlo simulation	[−0.183, 0.060]	[−0.175, 0.013]
Naive interval calculation	[−0.313, 0.115]	[−0.279, 0.093]
<i>Puffy</i>	Standardized skewness	Pearson's skewness
Monte Carlo simulation	[−1.654, 0.675]	[−0.778, 0.704]
Naive interval calculation	[−34.725, 8.981]	[−3.489, 3.019]

The naive interval calculations—which are so called pointedly because they do not account for the repeated parameters—yield outer estimates that are sure to enclose the true ranges of skewness but are probably too wide. The Monte Carlo simulations yield inner estimates that are almost certain to underestimate the true ranges. The discrepancies between these two kinds of bounding estimates reveals the computational consequences of not having exact methods for computing skewness as we do for means and variances.

## 4.8 Confidence intervals

Frequentist confidence intervals\* are widely used to characterize the sampling uncertainty of statistics computed from random sample data, i.e., from subsets of a

---

\*Confidence intervals might seem similar in spirit to the intervals considered in this report as data, but the similarity is only apparent and not profound in any way. There are two main differences. The first difference is that a confidence interval is a *statistical statement* about the data and the coverage one can expect for the parameter. It is not a sure claim with full confidence like the intervals the report has been discussing (see page 17). It is not asserting that the parameter is within the given interval. It is not even asserting anything about the probability that the parameter is within the interval. It is only characterizing the long-term performance in terms of coverage (by saying it is  $P\%$ ). The calculations made with intervals elsewhere in this report were based on a full-confidence assumption. Because confidence intervals lack this assumption, it is not mathematically reasonable to make calculations with confidence intervals as such. One could apply the rules of interval arithmetic (section 3.1) to confidence intervals, but the numerical results of such calculations would not generally have any particular meaning in terms of the coverage we might expect for the results in future experiments. Although it is sometimes possible to define joint confidence regions for bivariate or multivariate problems, the necessary mathematics is quite cumbersome and our abilities in such cases are entirely rudimentary.

The second main difference between the intervals considered in this report and traditional confidence intervals is that the latter are limited to characterizing sampling uncertainty that arises from not having measured every element in the statistical population. Our intervals, in contrast, can be used to characterize a variety of other kinds of uncertainty such as we described in section 2. This can include measurement uncertainty from intermittent or censored measurements, significant digits, data binning, privacy or security restrictions, missing data and gross ignorance. In principle, it might be possible to extend interval statistics as described in this report to intervals that also include sampling uncertainty as well as the other forms of uncertainty. This might be as simple as an analyst's appending an assumption of full confidence (as defined on page 17) to traditional confidence intervals computed at some confidence level that he or she considered to be good enough for a particular application. Grosz (1986), among many others, has suggested this. On the other hand, a mathematically correct treatment of sampling uncertainty within interval statistics that would be useful in engineering applications may require a more elaborate and careful integration of two epistemological views about data that are embodied in interval analysis and probabilistic sampling theory. This section on the application of interval statistics to computing confidence intervals is perhaps a first step in this direction, but a full discussion of this interesting subject is beyond the scope of this report.

population rather than the entire population. For instance, confidence intervals on the mean are very widely used as a characterization of the location of a data set, just like the mean, median, and mode are, except that confidence intervals also capture the effect of sample size (cf. Killeen 2005). They are wide when there are few samples and tighten as the sample size increases. In practice, confidence limits on means are often used as conservative estimates of data location when the paucity of samples makes the ordinary means unreliable estimates. It might be argued that confidence intervals are a part of inferential statistics, but we consider them here in their use as descriptions of the central tendency of a distribution.

A confidence interval for an unknown population parameter at the  $P\%$  level is a range defined in such a way that it will contain the true value of the parameter (at least)  $P\%$  of the time it is computed from random sample data. The value  $P\%$  is called the confidence level, or sometimes the coverage. As a result of this prescription, confidence intervals are not unique. There are many distinct confidence intervals that could be specified for a single parameter and a single data set that differ in how they are centered. The endpoints of confidence intervals are called upper and lower confidence limits, which can be symbolized by  $U$  and  $L$  respectively. The upper confidence limit is often of critical interest and is commonly denoted by the abbreviation UCL.

The most commonly used confidence limits are those for the mean of a variable assumed to follow a normal distribution. In this case, the confidence interval is

$$[L_{\alpha}, U_{\alpha}] = \left[ \bar{X} - t_{\alpha, N} \frac{s}{\sqrt{N}}, \bar{X} + t_{\alpha, N} \frac{s}{\sqrt{N}} \right]$$

where  $N$  is the sample size,  $\bar{X}$  is the sample mean,  $s$  is the sample standard deviation,  $t_{\alpha, N}$  is the critical value from the Student's  $t$  distribution associated with the desired confidence level and the given sample size. It was of course Gosset (Student 1908) who recognized the need for the multiplication by the  $t$  statistic to correct for the fact that the standard deviation was being estimated from data rather than known parametrically.

Computations of confidence limits usually ignore interval uncertainty in the data unless it has the form of non-detects (left censoring) which is widely recognized as problematic. In practice, many analysts pretend the uncertain values are zero or set them at the detection limit which is presumably the largest they could possibly be, or sometimes set them to a value that is one half the detection limit to split the difference. Some analysts actually use pseudorandom numbers to assign arbitrary values within the range between zero and the detection limit to each of the non-detects, and then compute the confidence limit as if those values had been measured. All of these “substitution” strategies are seriously flawed and cannot be used to estimate the confidence limit well or even conservatively (Helsel 1990; 2005). The problem is that the direction of bias of the substitution methods is unknown. Setting the non-detects to their detection limits will likely overestimate the mean, but will tend to underestimate the standard deviation. Conversely, setting non-detects to zero will underestimate the mean but overestimate the standard deviation. Because the confidence limits are functions of the *difference* between

the two statistics, we cannot be sure about whether they will be over- or underestimated. When censoring is moderate, Cohen's method is often used to account for non-detects (Gilbert 1987; EPA 2000, page 4-43ff). It is a maximum likelihood method for correcting the estimates of the sample mean and the sample variance to account for the presence of non-detects among the data. This method requires that the detection limit be the same for all the data and that the underlying data are normally distributed. Current guidance from the U.S. EPA (2002) suggests an interval statistics (bounding) approach to computing the bounds on the confidence limits in the presence of non-detects and other forms of data censoring using mathematical optimization, but it does not provide convenient algorithms for this.

Section 4.2.1 and 4.5.2 described ways to compute the interval average and interval standard deviation, and one might be tempted to just plug these intervals into the expression above. However, as often happens in interval computations, the resulting interval for  $U$  would be wider than the actual range because the values for the mean and the standard deviation are computed based on the same inputs  $x_1, \dots, x_N$  and cannot, therefore, change independently. Let us consider the simple case when  $x_1 = x_2 = [0, 1]$ . In this case, the range of the mean is  $(x_1+x_2)/2$  is equal to  $[0, 1]$ , where the largest value 1 is attained only if both values are 1. The sample standard deviation is  $[0, 1/2]$ , with the largest value attained only in two cases: when  $x_1 = 0$  and  $x_2 = 1$ , or when  $x_1 = 1$  and  $x_2 = 0$ . If we were simply to combine these intervals, we would conclude that  $U$  is within the interval  $[0, 1] + t_{95,2}[0, 1/2]/\sqrt{2} = [0, 1+t_{95,2}/\sqrt{8}]$ , where  $t_{95,2}$  is the critical value of the Student's  $t$  distribution that is associated with 95% confidence and a sample size of 2 (i.e., one degree of freedom). But it is easy to show that  $U$  cannot be as large as this interval suggests. The only way it could be this large is for both the mean and standard deviation to have their highest values. This is never possible with a single configuration of values for  $x_1$  and  $x_2$ , because the configuration that leads to the highest mean value doesn't lead to the highest value of the standard deviation and vice versa. Thus, the actual range of  $U$  must be narrower than would be computed by naively combining the two intervals.

The rest of section 4.8 concerns the derivation of algorithms to compute confidence limits and confidence intervals, both with and without an assumption of normality for the underlying distribution. If you don't need to know how this is done, you might skip to section 4.9 on page 75.

#### 4.8.1 Inner bounds on confidence limits

Kreinovich et al. (2003; 2004b; 2005a) described a feasible algorithm for computing the lower bound on the upper confidence limit  $\underline{U}$  and the upper bound on the lower confidence limit  $\bar{L}$ . These are the inner bounds on the endpoints of the confidence interval. The idea underlying this algorithm is similar to that for computing the lower bound on variance  $\underline{V}$ . The minimum of a differentiable function of a scalar value  $X_i$  over an interval is attained either inside this interval or at one of the endpoints. If the minimum is attained inside the interval, the derivative  $\partial U/\partial X_i$  is equal to zero. If function is minimized at the left endpoint  $X_i = \underline{x}_i$ , then  $\partial U/\partial X_i \geq 0$ , and if it is minimized at the right endpoint  $X_i = \bar{x}_i$  then  $\partial U/\partial X_i \leq 0$ . For the upper bound  $U$ , we have

$$\frac{\partial U}{\partial X_i} = \frac{1}{N} + K_0 \frac{X_i - E}{SN}$$

where  $K_0$  is some constant involving the critical value from the  $t$  distribution. So the derivative equals zero when and only when  $X_i$  takes on the value  $M = E - S/K_0$ . Comparing  $M$  with  $X_i$  likewise tells us whether the derivative is non-positive or non-negative, so we can say that either

$$\begin{aligned} M &\leq X_i = \underline{x}_i, \\ X_i &= \bar{x}_i \leq M, \text{ or} \\ X_i &= M \in ]\underline{x}_i, \bar{x}_i[ , \end{aligned}$$

(where the inverted brackets denote the open interval that excludes its endpoints). Hence, if we know the location of the value  $M$  with respect to all the intervals  $x_i$ , we will immediately know the values of the scalar configuration  $X_i$  that extremizes  $U$  by these three rules. To find the minimum, therefore, we will analyze how the endpoints of the input intervals divide the real line and consider all the resulting zones.

Let us sort all  $2N$  endpoints together into one sequence and denote them by  $Y_1 \leq \dots \leq Y_{2N}$ . For each input interval  $i$  for which  $M$  is not an element of the interior of the interval, the scalar value  $X_i$  that extremizes  $U$  is uniquely determined by the first two rules above. For each input interval  $i$  for which  $M \in ]\underline{x}_i, \bar{x}_i[$ , the extremizing scalar value  $X_i$  should be equal to  $M$ , and that  $M$  must be the same for all such intervals. To determine what it is, we use its definition  $M = E - S/K_0$  where the values of  $E$  and  $S$  are computed from the same configuration of scalars  $X_i$ . This equation is equivalent to  $0 \leq E - M$  and  $S^2/K_0^2 = (M - E)^2$ . Substituting the values of  $X_i$  into the formula for the mean  $E$  and for the standard deviation  $S$ , we get a quadratic equation for  $M$ . For each zone of the real line partitioned by the  $Y$ 's, we can uniquely determine the values  $X_i$  that *may* correspond to a minimum of  $U$ . These considerations justify the following algorithm.

*Algorithm for computing inner bounds on confidence limits  $\underline{U}$  and  $\bar{L}$ .*

0. Data are  $x_i = [\underline{x}_i, \bar{x}_i]$ ,  $i = 1, \dots, N$ .
1. Sort all  $2N$  endpoints  $\underline{x}_i, \bar{x}_i$  into a single sequence  $Y_1 \leq Y_2 \leq \dots \leq Y_{2N}$ . Let  $Y_0 = -\infty$  and  $Y_{2N+1} = +\infty$ . The real line is thus partitioned into  $2N+1$  zones.
2. For each zone  $[Y_k, Y_{k+1}]$ ,  $k = 0, 1, \dots, 2N$ , do the following steps:
  - a. Compute the values

$$\begin{aligned} e_k &= \sum_{i: \bar{x}_i \leq Y_k} \bar{x}_i + \sum_{j: \underline{x}_j \geq Y_{k+1}} \underline{x}_j \\ m_k &= \sum_{i: \bar{x}_i \leq Y_k} \bar{x}_i^2 + \sum_{j: \underline{x}_j \geq Y_{k+1}} \underline{x}_j^2, \end{aligned}$$

and let  $N_k$  be the total number of such  $i$ 's and  $j$ 's.

- b. Letting  $\alpha = 1/K_0$ , compute the values

$$\begin{aligned} A_k &= e_k^2 (1 + \alpha^2) - N \alpha^2 m_k, \\ B_k &= 2 e_k (N_k (1 + \alpha^2) - N \alpha^2), \text{ and} \\ C_k &= N_k (N_k (1 + \alpha^2) - N \alpha^2). \end{aligned}$$

- c. Solve the quadratic equation  $A_k - B_k M + M^2 = 0$  to obtain  $M$ .
- d. For computing  $\underline{U}$ , select only those solutions for which  $M N_k \leq e_k$  and  $M \in [Y_k, Y_{k+1}]$ . For computing  $\bar{L}$ , select only those solutions for which  $M N_k \geq e_k$  and  $M \in [Y_k, Y_{k+1}]$ . For each selected solution, compute the values of

$$\begin{aligned} E_k &= \frac{e_k}{N} + \frac{N - N_k}{N} M, \\ W_k &= \frac{m_k}{N} + \frac{N - N_k}{N} M^2, \\ U_k &= E_k + K_0 \sqrt{W_k - E_k^2}, \\ L_k &= E_k - K_0 \sqrt{W_k - E_k^2}. \end{aligned}$$

3. The smallest  $U_k$  computed in step 2d is the desired lower bound on  $U$ . The largest  $L_k$  computed in step 2d is the desired upper bound on  $L$ .

This algorithm is fairly efficient. Sorting the endpoints requires  $\Theta(N \log N)$  steps (Cormen 2001). The quantities  $e_k$ ,  $m_k$  and  $N_k$  computed in step 2a at any step  $k > 1$  depend simply on the analogous values computed in the previous step  $k-1$ . We can improve the algorithm in an obvious way so that its initial computation of all quantities requires linear time  $\Theta(N)$ , and for each zone, it needs only a constant-time operation to update all quantities, solve the quadratic equation, and compute the corresponding values of the confidence limits. Thus, the algorithm requires  $\Theta(N \log N)$  time overall.

*Numerical example.* Suppose that  $K_0 = 2$  and we have  $N = 2$  intervals,  $[-2, -1]$  and  $[1, 2]$  and we want to compute the upper bound on the upper confidence limit. In this example, sorting leads to  $Y_1 = -2$ ,  $Y_2 = -1$ ,  $Y_3 = 1$ ,  $Y_4 = 2$ , so the real line is divided into five zones,  $[-\infty, -2]$ ,  $[-2, -1]$ ,  $[-1, 1]$ ,  $[1, 2]$ , and  $[2, \infty]$ . We show the computations for each of these five zones. In this case,  $[Y_0, Y_1] = [-\infty, -2]$ . Because  $\bar{x}_j \geq -2$  for all  $j$ , the sum  $e_0$  gets the value  $(-2) + 1 = -1$ , and  $m_0$  gets the value  $(-2)^2 + 1^2 = 5$ , and  $N_0$  is 2. Because  $\alpha = 1/K_0 = 1/2$ , we have

$$\begin{aligned} A_0 &= (-1)^2 \times (1 + (1/2)^2) - 2 \times (1/2)^2 \times 5 = -5/4, \\ B_0 &= 2 \times (-1) \times (2 \times (1 + (1/2)^2) - 2 \times (1/2)^2) = -4, \text{ and} \\ C_0 &= 2 \times (2 \times (1 + (1/2)^2) - 2 \times (1/2)^2) = 4. \end{aligned}$$

The corresponding quadratic equation  $4M^2 + 4M - 5/4 = 0$  has two roots

$$M_{1,2} = \frac{-4 \pm \sqrt{4^2 + 4 \times (5/4) \times 4}}{2 \times 4} = \frac{-4 \pm 6}{8}.$$

Neither of the roots,  $M_1 = -1.25$  or  $M_2 = 0.25$ , is in the zone  $[-\infty, -2]$ . For the second zone,  $[Y_1, Y_2] = [-2, -1]$ ,  $e_1 = 1$ ,  $m_1 = 1$ , and  $N_1 = 1$ . Consequently,  $A_1 = 3/4$ ,  $B_1 = 6/4$ , and  $C_1 = 3/4$ . Diving both sides of the quadratic equation by  $3/4$  yields  $M^2 - 2M + 1 = 0$ , whose double root  $M_{1,2} = 1$  is outside the zone. For the third zone  $[-1, 1]$ ,  $e_2 = 0$ ,  $m_2 = 2$ , and  $N_2 = 2$ , so  $A_2 = -1$ ,  $B_2 = 0$ , and  $C_2 = 4$ , and the roots  $M_{1,2} = \pm 1/2$ , both of which are inside the zone. However, only the root  $-1/2$  satisfies the inequality  $M N_2 \leq e_2 = 0$ . For this root,  $E_2 = 0/2 + (2 - 2)/2$

$$E_2 = \frac{0}{2} + \frac{2-2}{2} \times \left(-\frac{1}{2}\right) = 0,$$

$$W_2 = \frac{2}{2} + \frac{2-2}{2} \times \left(-\frac{1}{2}\right)^2 = 1,$$

and

$$U_2 = 0 + 2\sqrt{1 - 0^2} = 2.$$

For the fourth zone,  $[1, 2]$ ,  $e_3 = -1$ ,  $m_3 = 1$ , and  $N_3 = 1$ , so  $A_3 = 3/4$ ,  $B_3 = -6/4$ , and  $C_3 = 3/4$ . Again, diving both sides of the quadratic equation by  $3/4$  yields  $M^2 + 2M + 1 = 0$ , whose double root  $M_{1,2} = -1$  is outside the zone. Finally, for the fifth zone  $[2, \infty]$ , all of the right endpoints of the input intervals are less than or equal to 2, so  $e_4 = 1$ ,  $m_4 = 5$ , and  $N_4 = 2$ , so  $A_4 = -5/4$ ,  $B_4 = 4$ , and  $C_4 = 4$ . The corresponding quadratic equation has roots  $M_{1,2} = (4 \pm 6)/8$ , both of which are outside the zone. The largest (and only) value of  $U$  computed was 2, so the upper bound on the upper confidence limit on the mean is 2.

#### 4.8.2 Computing the outer bounds on confidence limits

Computing the upper bound on the upper confidence limit  $\bar{U}$  or computing the lower bound on the lower confidence limit  $\underline{L}$  is an NP-hard problem (Kreinovich et al. 2003; 2004b; 2005a). If the quantity  $\beta = 1 + 1/K_0^2 \leq N$ , then the corresponding maximum and minimum are always attained at the endpoints of the intervals  $x_i$ , so to compute the outer bounds on the confidence limits it is sufficient to consider all  $2^N$  combinations of such endpoints, which are the corners of the Cartesian product of the input intervals. This is true, for instance, if  $K_0$  is larger than one and the sample size  $N$  is two or larger, in other words, for almost all cases of practical interest. The sections below consider some special cases of interval data that offer better prospects computationally than this general case.

*Numerical example.* Let us compute upper and lower bounds on the upper confidence limit on the mean (UCL) for the skinny and puffy data sets. The skinny and puffy data were given on page 28 and plotted in Figure 6. Assuming their underlying distributions are normal, we can use the corner-checking algorithm just described to compute the upper bounds on the 95% UCLs and the algorithm described in section 4.8.1 to compute

the lower bounds. These calculations are a bit too cumbersome to reproduce here; for instance, the corner-checking algorithm for the puffy data requires calculating values for each of  $2^9 = 512$  configurations of scalars from within the respective intervals. We used one-sided critical  $t$ -values. The resulting normal one-sided 95% UCL is  $[5.958, 8.863]$  for the puffy data and  $[8.070, 8.613]$  for the skinny data, both with outward-directed rounding. These interval UCLs are plotted in Figure 17. Note that UCL for the skinny data has a smaller upper bound than that for the puffy data, even though it was based on fewer data. Keep in mind that the intervals displayed in the figure are *not* analogous to confidence intervals for ordinary scalar data sets, but rather to the upper confidence *limit*, which is always a scalar value whenever the data are all scalars.

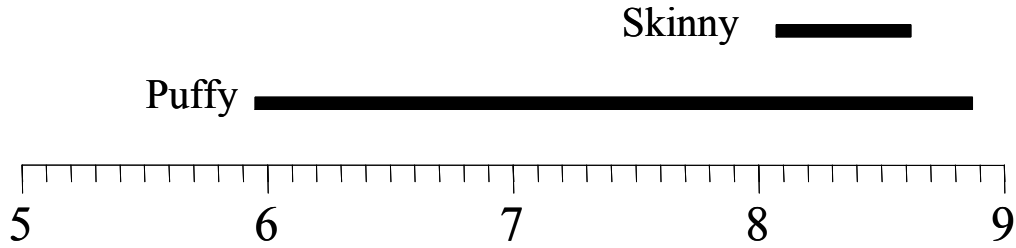


Figure 17. Normal one-sided 95% upper confidence limits on the mean (UCLs) for the skinny and puffy data.

#### 4.8.2.1 No-intersection and scattered data

The outer confidence limits can be calculated efficiently in the special case when the interval data  $x_i, i = 1, \dots, N$ , do not intersect with each other, or when they are scattered in the sense that the narrowed intervals

$$y_i = \left[ \tilde{x}_i - \beta \frac{\Delta_i}{N}, \tilde{x}_i + \beta \frac{\Delta_i}{N} \right] = \left[ \underline{x}_i + \bar{x}_i - \beta \frac{(\bar{x}_i - \underline{x}_i)}{N}, \underline{x}_i + \bar{x}_i + \beta \frac{(\bar{x}_i - \underline{x}_i)}{N} \right] / 2,$$

do not intersect, where  $i = 1, \dots, N$ ,  $\beta = 1 + 1/K_0^2$ , and  $\tilde{x}$  and  $\Delta$  denotes the midpoint and halfwidth of an interval. These narrowed intervals have the same midpoints as the original data  $x_i$ , but they can have much smaller widths, depending on the sample size  $N$  and  $K_0$ . In fact, a feasible algorithm is even possible when no more than  $K < N$  of the narrowed intervals have a common intersection (Kreinovich 2005a).

*Algorithm for computing outer bounds on confidence limits  $\underline{L}$  and  $\bar{U}$  for scattered data.*

0. Data are  $x_i = [\underline{x}_i, \bar{x}_i]$ ,  $i = 1, \dots, N$ , such that, given  $\beta = 1 + 1/K_0^2$ , the narrowed intervals

$$y_i = \left[ \underline{x}_i + \bar{x}_i - \beta \frac{(\bar{x}_i - \underline{x}_i)}{N}, \underline{x}_i + \bar{x}_i + \beta \frac{(\bar{x}_i - \underline{x}_i)}{N} \right] / 2$$

have no mutual intersection.



1. Sort all  $2N$  endpoints of the narrowed intervals  $y_i, \bar{y}_i$  into a single sequence  $Y_1 \leq Y_2 \leq \dots \leq Y_{2N}$ . Let  $Y_0 = -\infty$  and  $Y_{2N+1} = +\infty$ . The real line is thus partitioned into  $2N+1$  zones.
2. For each zone  $[Y_k, Y_{k+1}]$ ,  $k = 0, 1, \dots, 2N$ , and for each  $j$  from 1 to  $N$ , select one or more values for  $X_j$  according to these rules:
  - a. If  $Y_{k+1} < y_j$ , then set  $X_j = \bar{x}_j$ .
  - b. If  $Y_{k+1} > y_j$ , then set  $X_j = \underline{x}_j$ .
  - c. For all other  $j$ , set  $X_j$  to both possible values  $\bar{x}_j$  and  $\underline{x}_j$  and consider both sequences.

The result will be one or multiple configurations of  $X_j$  for each zone.

3. For each of the configurations  $X_j$  assembled in step 2,
  - a. Compute the mean  $E$  and standard deviation  $S$  for the configuration  $X_j$ .
  - b. If  $E - S/K_0 \in [Y_k, Y_{k+1}]$ , then compute  $U = E + K_0 S$ .
  - c. If  $E + S/K_0 \in [Y_k, Y_{k+1}]$ , then compute  $L = E - K_0 S$ .
4. The largest  $U$  computed in step 3b is the desired upper bound on the upper confidence limit on the mean  $\bar{U}$ . The smallest  $L$  computed in step 3c is the desired lower bound for the lower confidence limit on the mean  $\underline{L}$ .

These algorithms require quadratic time. However, modifications to the algorithms are possible that would enable the computations to be done in only  $\mathcal{O}(N \log N)$  time. The modifications are similar to those described in section 4.5.1.2.6 on page 55 for computing the upper bound on variance (see also Xiang 2006).

#### 4.8.2.2 No-nesting data

For the case of no-nesting data, we can prove that the maximum of  $U$  and the minimum of  $L$  are attained at one of the corners of the Cartesian product of the input intervals. In other words, the configuration of scalar values that corresponds to these extremes consists of either the right or left endpoint from every interval. This is similar to what we saw for the problem of computing the upper bound on variance in section 4.5.1.2.2. Indeed, practically the same proof works, because increasing  $V$  without changing  $E$  will also increase  $U = E + K_0 \sqrt{V}$  as well.

Therefore, an efficient algorithm would first sort the intervals  $x_i$  into lexicographic order so that  $x_i \leq x_j$  if and only if either  $\underline{x}_i \leq \underline{x}_j$ , or both  $\underline{x}_i = \underline{x}_j$  and  $\bar{x}_i \leq \bar{x}_j$ . Then, for each  $k = 0, 1, \dots, N$ , the algorithm would compute the values  $V = M - E^2$  and  $L$  and  $U$  for the corresponding configurations  $x^{(k)} = (\underline{x}_1, \dots, \underline{x}_k, \bar{x}_{k+1}, \dots, \bar{x}_N)$ . As in the algorithm for variance, when we go from a vector  $x^{(k)}$  to the vector  $x^{(k+1)}$ , only one component changes, so only one term changes in each of the sums  $E$  and  $M$ , which makes them easy and fast to compute sequentially.

Sorting takes  $\mathcal{O}(N \log N)$  time. Computing the initial values of  $E$  and  $M$  requires linear time  $\mathcal{O}(N)$ . For each  $k$ , computing the new values of  $E$  and  $M$  requires a constant number of steps so, overall, computing all  $N$  values of  $E$ ,  $M$  (and hence  $V$ ) requires linear time. Thus, the overall time of this algorithm is  $\mathcal{O}(N \log N)$ .

#### **4.8.2.3 Arrangeable data**

Outer confidence limits can be computed for arrangeable data with an algorithm similar to that used for computing the upper bound of variance (section 4.5.1.2.3 on page 51). This allows computation for this special case in  $\mathcal{O}(N^D)$  where  $D$  is the number of subgroups in the data set that each have no nesting.

### **4.8.3 Confidence intervals without a normality assumption**

This section considers the problem of computing confidence intervals about the mean when the normality assumption about the underlying distribution is not tenable.

#### **4.8.3.1 Non-normal distributions**

The confidence limits described above assume that the underlying data are drawn from a normal distribution, which may not always be a reasonable assumption. When the underlying population cannot be assumed to be normal, the central limit theorem is sometimes employed to estimate approximate confidence limits on the mean, provided that  $N$  is reasonably large. This method can easily be generalized to handle interval data. However, the method is rather sensitive to skewness in the underlying distribution and cannot be recommended for skewed data (Johnson 1978; Chen 1995; Zhou and Gao 2000; EPA 2002). Specific methods for a few specific distributions have been described, including the gamma distribution (Wong 1993; Stuart and Ord 1994) and the lognormal distribution (Land 1972; 1975; Gilbert 1987; Zhou and Gao 1997), but these cannot readily be generalized for interval data because they use complex expressions based on maximum likelihood. Alternative methods (e.g., Hall 1988; 1992; Schulz and Griffin 1999) can also be complicated, and the basic problems of skewness and uncertainty about the shape of the underlying distribution—even before the issue of censoring or interval uncertainty—make this area of statistics somewhat controversial.

#### **4.8.3.2 Distribution-free confidence limits on the mean**

We can also compute distribution-free confidence limits on the mean, which make no assumption about the distribution shape, by using the Kolmogorov-Smirnov confidence limits on the distribution of the data (section 4.1.2). Because these are distribution-free bounds on the distribution, the means computed from the left and right sides of the p-box therefore represent distribution-free bounds on the mean. This construction assumes only that the data are randomly sampled and the underlying distribution is stationary and continuous. For the limits on the mean to be finite, however, this method requires known lower and upper finite constraints on the underlying variable. Although this approach is probably very conservative if the constraints on the variable are loose, it represents the only\* known method for obtaining true confidence limits on the mean that does not depend on an analyst's making shape assumptions about the distribution of the variable.

---

\*Contra Estler 1997; contra Singh et al. 1997; 2000; Schulz and Griffin 1999; see Huber 2001.

*Numerical example.* Kolmogorov-Smirnov 95% confidence bands on the distributions for the skinny and puffy data distributions were shown in Figure 8. We can compute distribution-free confidence intervals on the mean for these data sets from the bounds of the Kolmogorov-Smirnov confidence bands. If we truncate these distributions at zero and twenty—saying the values of  $X$  cannot be outside this range—then the confidence intervals on the mean will be finite. The resulting distribution-free 95% confidence intervals on the mean for the skinny and puffy data are [1.160, 14.694] and [1.706, 13.787], respectively, and are depicted in Figure 18. Comparing this figure with Figure 17, we may note that, without the normality assumption, the right bound for the skinny data is now larger than the right bound for the puffy data. (Keep in mind, however, that the confidence *intervals* in Figure 18 are not exactly analogous to the upper confidence *limits* shown in Figure 17.)

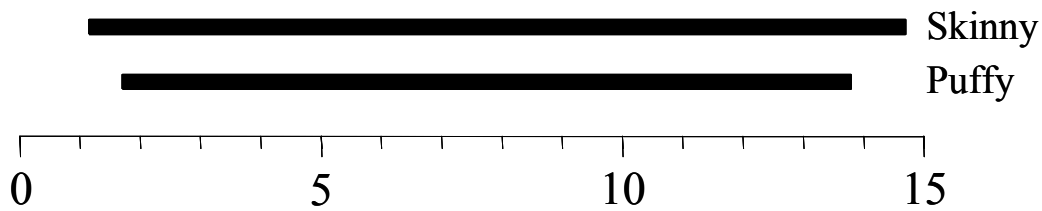


Figure 18. Distribution-free 95% confidence intervals on the mean for the skinny and puffy data.

#### 4.8.4 Confidence limits for other statistics

Horowitz and Manski (2000) considered the problem of obtaining a joint confidence interval for a pair of bounds with asymptotic probability of containing both bounds on an interval population parameter. They focused on confidence intervals of the form

$$[\underline{s}_n - Z_{n\alpha}, \bar{s}_n + Z_{n\alpha}]$$

where  $\underline{s}_n$  and  $\bar{s}_n$  are the respective sample estimates based on  $n$  samples of the interval bounds  $\underline{s}$  and  $\bar{s}$  on the population parameter of interest, and  $Z_{n\alpha}$  is chosen so that asymptotically  $P(\underline{s}_n - Z_{n\alpha} \leq \underline{s}, \bar{s} \leq \bar{s}_n + Z_{n\alpha}) = 1 - \alpha$ . They suggested deriving  $Z_{n\alpha}$  from the asymptotic bivariate distribution of  $\underline{s}_n$  and  $\bar{s}_n$ , or obtaining it from a bootstrap simulation.

#### 4.9 Fitting named distributions to interval data

Analysts often need to fit probability distributions from particular families (like normal, lognormal, Weibull, etc.) to sample data. For instance, the standard approach to modeling measurement uncertainty (to be discussed in section 8) fits data to normal distributions. There are several common strategies for fitting named distributions to data, including the method of matching moments, regression techniques, maximum likelihood

methods and kernel smoothing methods. A thorough discussion of fitting distributions to interval data sets is really beyond the scope of this report, but this section outlines the main approaches and the prospects for developing necessary algorithms to extend them to the case of data containing incertitude.

#### 4.9.1 Method of matching moments

Perhaps the most commonly used approach for fitting distributions to data sets is the method of matching moments, in which the distribution is assumed to have the same moments about the origin as the observed data. This method of fitting distributions to data is very easy to apply for distribution shapes that can be specified by a single parameter, such as the Bernoulli, binomial (if the number of trials is fixed), chi-squared, exponential, geometric, Pascal, Poisson, Rayleigh, and reciprocal distributions. For these shapes, only the mean of the data needs to be computed to define the fitted distribution. If this mean is an interval rather than a point estimate, then fitting by the method of matching moments yields a p-box rather than a precise distribution. Ferson et al. (2003) described how to derive the p-box specified by a single interval parameter. For each of the distributions named above, the fitted p-box is simply the envelope of the two distributions corresponding to the two endpoints of the mean. The resulting p-box encloses all distributions of the specified shape that could be fitted by the method of moments to a configuration of scalar values taken from within the respective intervals in the data set. It can be said that the p-box represents the assumption about the shape of the underlying distribution, as informed by the available data.

*Numerical examples.* Suppose we want to fit exponential distributions to the skinny and puffy data (not that these data particularly suggest this would be a good model). As we saw in section 4.2.1, the means of the skinny and puffy data sets are [5.338, 5.849] and [4.56, 7.39] respectively. If exponential distributions are parameterized by  $\lambda$  such that  $1/\lambda$  is the distribution mean, then  $\lambda$  for the skinny data set is [0.1709, 0.1874] and  $\lambda$  for the puffy data set is [0.135, 0.220]. The classes of all exponential distributions that have means inside these ranges respectively are depicted as the p-boxes shown in Figure 19.

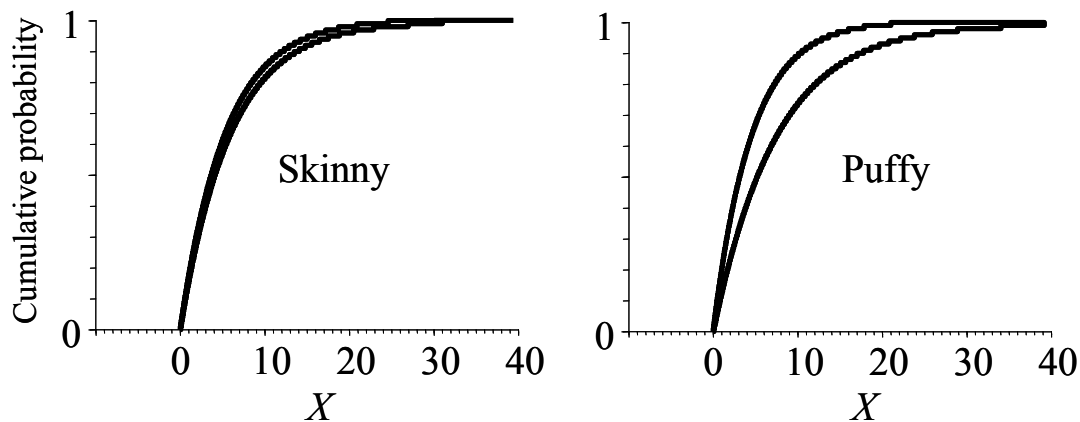


Figure 19. P-boxes from fitting exponential distributions to the skinny and puffy data sets by the method of matching moments.

The specifications of many other distributions require two parameters, such as the beta, Erlang, extreme value, gamma, Gaussian, Gumbel, Laplace, logistic, lognormal, loguniform, normal, Pareto, power function, rectangular, and uniform distributions. For these shapes, both the mean and the second moment need to be computed from the data set to define the fitted distribution. The approach described in Ferson et al. (2003) extends to this two-parameter case, and it produces a p-box that is guaranteed to enclose the named distribution fitted to the interval data by the method of matching moments. For example, the parameters of a beta distribution are

$$v = E(x) \left( \frac{E(x)(1 - E(x))}{E(x^2) - E(x)} - 1 \right)$$

$$w = (1 - E(x)) \left( \frac{E(x)(1 - E(x))}{E(x^2) - E(x)} - 1 \right)$$

where  $E(x)$  is the arithmetic mean of the data and  $E(x^2)$  is their second moment about the origin. But when the input data are intervals rather than point estimates, the appearance of repeated parameters in these expressions suggests that using naive interval analysis to compute them will overstate the uncertainty in the result. Not only is  $E(x)$  repeated; both of the terms  $E(x)$  and  $E(x^2)$  implicitly contain each of the underlying interval data, so that means that each value  $x_i$  is in fact repeated several times in both expressions. If any of these values are intervals, their uncertainties are inserted into the evaluations multiple times. Even if we computed best possible bounds on  $v$  and  $w$  with special strategies such as subinterval reconstitution (Moore 1966; Corliss 1988; Ferson and Hajagos 2004), there would still be a problem because the resulting two interval parameters are not independent, and the p-box constructed from them would not be the best possible p-box given the sample data themselves. The reason is that the moments are not independent of each other (and any parameters derived from them would also not be independent).

Understanding the source of the dependence between the interval parameters suggests a strategy around the difficulty. Consider the bounds on the mean and the standard deviation. The largest possible mean cannot be paired with the largest possible standard deviation because there is no configuration of scalar values within their respective intervals of the data set that extremize both the mean and the standard deviation. A strategy that yields a much tighter result can be derived by taking account of this dependence.

Let us consider the problem for the normal distribution as an important special case. The cumulative distribution function for any normally distributed variable  $N(\mu, \sigma)$  with mean  $\mu$  and standard deviation  $\sigma$  is

$$F(X) = \Phi\left(\frac{X - \mu}{\sigma}\right)$$

where  $\Phi$  is the standard normal distribution function\* which is a strictly increasing function of  $X$ . This means that, for every value of  $X$ , the interval bounds on the distribution function can be computed as  $[\Phi(\underline{r}), \Phi(\bar{r})]$ , where  $[\underline{r}, \bar{r}]$  is the interval of possible values of the ratio  $r = (X - \mu) / \sigma$ . Thus, to find the bounds on the normal distributions at any value of  $X$ , it is sufficient to find this interval  $[\underline{r}, \bar{r}]$ .

Suppose we fix the value along the  $X$ -axis for which we'll compute the vertical bounds on the distribution functions. The value of  $r$  varies as a continuous function of the configuration of scalar values from within the Cartesian product of input intervals  $x_1 \times x_2 \times \dots \times x_N$ , so its smallest and largest values are attained at some configurations (although  $r$  may be infinite if  $\sigma$  is zero). Let  $(X_1^*, X_2^*, \dots, X_N^*)$  be a configuration at which  $r$  is maximized. Consider a function  $r(X_1^*, \dots, X_i, \dots, X_N^*)$  whose arguments are all fixed at these extremizing points except for the  $i$ th one, which we might vary. (Of course when  $X_i = X_i^*$  the function is maximal.) There are three possible cases:

- If the ratio  $r$  attains its maximum at  $X_i \in ] \underline{x}_i, \bar{x}_i [$ , i.e., *inside the data interval*, then by calculus  $\partial r / \partial X_i = 0$  at this point.
- If  $r$  attains its maximum at the left endpoint  $X_i = \underline{x}_i$ , then the derivative  $\partial r / \partial X_i$  at this point cannot be positive. If it were, we would have even larger values of  $r$  for  $X_i$  larger than  $\underline{x}_i$ .
- Similarly, if  $r$  attains its maximum at the right endpoint, then the derivative must be larger than or equal to zero.

Analogous considerations tell us about the derivatives at the minimum for  $r$ . Analysis of these elementary considerations leads to the following algorithm (Xiang et al. 2007b).

*Algorithm for bounding normal distributions fitted by the method of moments.*

0. Data are  $x_i = [ \underline{x}_i, \bar{x}_i ]$ ,  $i = 1, \dots, N$ , which are assumed to be randomly sampled from a normal distribution.
1. Select a scalar value  $X$  at which bounds on the fitted cumulative distribution are to be evaluated.
2. For every possible partition of the indices  $\{1, \dots, N\}$  into the sets  $I^-$ ,  $I^+$ , and  $I^0$ , do:
  - a. For every  $i \in I^-$ , set  $X_i = \underline{x}_i$ .
  - b. For every  $i \in I^+$ , set  $X_i = \bar{x}_i$ .
  - c. If  $I^0$  is not empty, then:

---

\*The function symbolized as  $\Phi(X)$  is the integral from  $-\infty$  to  $X$  of the standard normal density function  $\phi(X) = (2\pi)^{-1/2} \exp(X^2/2)$ . It has no closed-form expression, but it is widely available in computer implementations. For instance, in Excel spreadsheets it can be evaluated with the formula NORMSDIST( $X$ ). In the R statistical programming environment, it can be computed with the formula pnorm( $X$ ).

i. Compute

$$A = \sum_{i \in I^-} x_i + \sum_{j \in I^+} \bar{x}_j = \sum_{k \in I^- \cup I^+} X_k, \text{ and}$$

$$M = \sum_{i \in I^-} (x_i)^2 + \sum_{j \in I^+} (\bar{x}_j)^2 = \sum_{k \in I^- \cup I^+} (X_k)^2.$$

ii. Let  $N_0$  denote the cardinality of  $I^0$ , and solve the quadratic equation

$$M + \frac{1}{N_0} (NB - A)^2 = N \left( \frac{NB - A}{N_0} \right) (x - B) + NB^2$$

for the value of  $B$ .

iii. If the quadratic equation has no real solutions, skip this partition and go to the next one.

iv. If at least one of the real solutions of  $B$  belongs to all the intervals  $x_i, i \in I^0$ , then set  $X_i = (NB - A) / N_0$ .

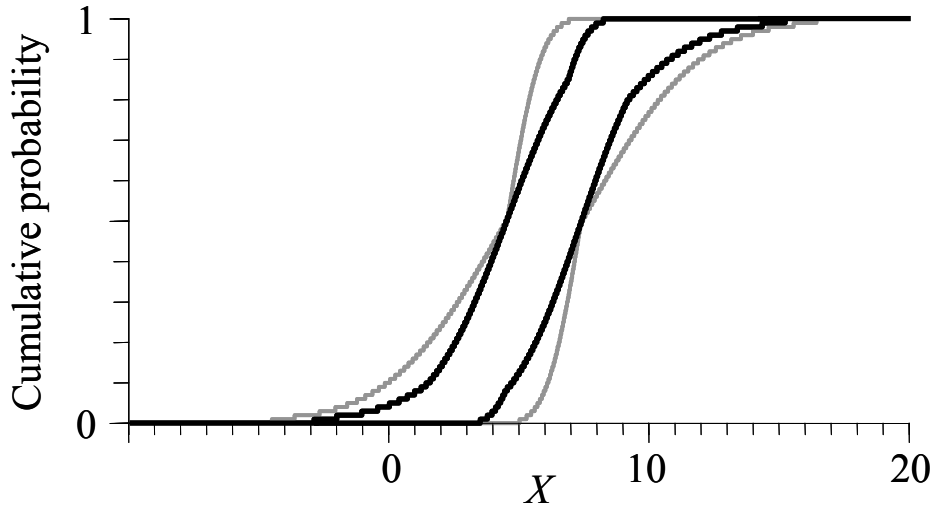
d. Compute  $r$  for every non-skipped partition.

3. Let  $\underline{r}$  be the smallest value of  $r$  computed in step 2d, and let  $\bar{r}$  be the largest value of  $r$  computed in step 2d.
4. The bounds  $[\Phi(\underline{r}), \Phi(\bar{r})]$ , where  $\Phi$  is the standard normal distribution function, are the desired bounds on the fitted distributions at the value of  $X$  selected in step 1.
5. Repeat steps 1 through 4 for all values of  $X$  for which bounds on the fitted distribution are sought.

Xiang et al. (2007b) show this algorithm always produces exact (i.e., both rigorous and best-possible) bounds on the distributions that would be fitted by the method of moments to scalar data consistent with the given intervals, but it uses a rather brute-force approach to solving the problem. For each of the  $N$  indices, we have three choices (assigning to sets  $I^-, I^+$ , or  $I^0$ ). That means there are  $3^N$  possible assignments. Thus this algorithm requires an exponential number of computational steps. (Although the algorithm requires a separate search for each  $X$ -value along a continuum of such values, in practice, we would pick some finite number of values, say  $N_p$ . There would thus be  $N_p$ -times more number of steps, but this factor doesn't grow with  $N$  so it is neglected in the order calculation, even though it may need to be rather large.) A much more efficient algorithm to bound normal distributions fitted to interval data could probably be devised, but it awaits discovery.

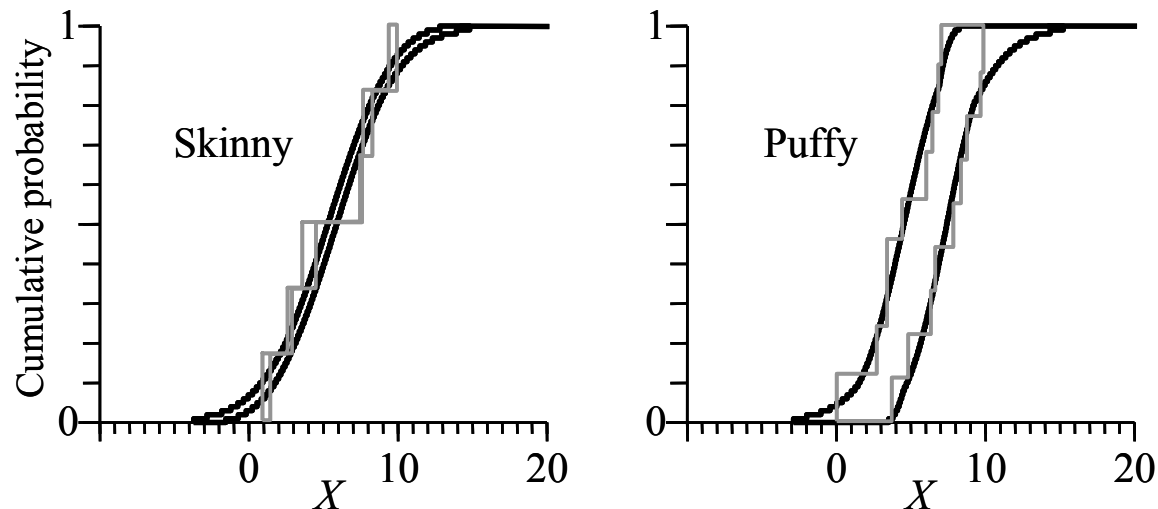
*Numerical example.* Suppose we assume the underlying distribution for the puffy data set is normal and seek to fit bounds on this normal distribution by the method of matching moments. As we saw in section 4.6.1, the first and second moments of the puffy data set are  $[4.56, 7.39]$  and  $[25.56, 58.54]$  respectively. These correspond to a mean and variance of  $[4.56, 7.39]$  and  $[0.9164, 10.9745]$  (which is the population variance, rather than the sample variance). The p-box formed naively from these interval

parameters for a normal distribution is depicted in Figure 20 as the outer bounds (shown in thin gray lines). The tighter black p-box represents the best possible bounds on a normal distribution fitted to the 9 interval values in the puffy data set by the method of matching moments using the algorithm above with  $3^9 = 19,683$  steps for each value of  $X$ . It evaluated vertical bounds at  $N_p = 100$  values along the  $X$ -axis, and used an outward-directed discretization scheme (Ferson and Hajagos 2004; section 3.5.6.3 of Ferson et al. 2003) for plotting to ensure the bounds retain their rigorousness for intermediate values of  $X$  for which exact bounds were not computed. This outward-directed plotting creates the step-function appearance which may be visible in the graph of the bounds. The breadth of this p-box is the result of the incertitude in these data. The enclosing gray p-box is also guaranteed to bound the fitted normal distributions, but does not do so optimally because it does not take account of the dependencies among the moments of possible configurations of scalar values within their respective intervals. To compare the fit against the data themselves, Figure 21 shows the fitted p-boxes in black, for both the skinny and puffy data sets, and superimposes in gray the empirical distributions originally depicted in Figure 6.



**Figure 20. P-box (black) for normal distributions fitted to the puffy data set by the method of matching moments and an enclosing p-box (gray) fitted to the bounds on the moments only.**





**Figure 21. P-boxes (black) for normal distributions fitted by the method of matching moments and the corresponding empirical distributions (gray).**

Some distributions, such as Simpson's triangular distribution and the trapezoidal distribution require three, four or more parameters, so would have to match in the mean, variance, and skewness, kurtosis, etc. The naive approach of computing a p-box based on three or more interval moments without accounting for the dependencies among the moments would, again, be guaranteed to enclose the true distribution, but would not yield the best possible p-box that could do so.

The method of matching moments is not always available. For instance, it cannot be used for the Cauchy distribution whose moments are infinite. Although the method of moments is far-and-away the most commonly used approach for fitting distributions to data, it lacks certain desirable optimality properties enjoyed by other approaches such as maximum likelihood and regression techniques.

The rest of section 4.9 reviews other ways to fit named distributions to data sets containing intervals. You may elect to skip to section 5 on page 88.

#### 4.9.2 Fitting distributions by regression

Besides the method of moments, another commonly used technique for fitting named distributions to a data set is an approach based on linear regression. In this approach, the data values are ranked, and the ranks are transformed to grades by dividing each rank by  $(N + 1)$ , so the grades range on the interval  $[0, 1]$ . The grades are then subjected to a probability transform associated with the distribution shape being fitted. The transformed grades and the data are subjected to linear regression analysis, and the statistics from the regression are interpreted to parameterize the fitted distribution. For the normal distribution, the transformation is the inverse cumulative distribution function for the standard (zero mean and unit variance) normal distribution. The original data values are then linearly regressed on their thus transformed grades. The intercept of the best fit line

is taken as the fitted mean of the normal distribution and the slope is taken as the standard deviation.

Although the method of fitting distributions by linear regression seems to require ranking the input data, we can extend this method for an interval data set. Although intervals cannot generally be ranked by magnitude, it is possible under the spike interpretation of intervals to rank the unknown scalar quantities represented by the intervals. The values associated with the ranks are, naturally, imprecise. Interval regression (e.g., Markov 1990; Manski and Tamer 2002; Marino and Palumbo 2002; 2003) can then be used to complete the distribution fitting. The regression technique used must allow intervals in the dependent variable. There are several different kinds of interval regressions that might be used, but this topic is beyond the scope of this report. Although the approach outlined here analogizes the use of regression techniques to fit a distribution to a scalar data set and generalizes it for interval data, it is not clear whether this approach can be structured to yield the best possible bounds on the family of distribution functions that would be obtained from considering all possible configurations of scalar points within their respective intervals of the data set and fitting distributions by regression to these configurations of scalar values.

*Numerical examples.* The approach is intuitively straightforward for the skinny data set displayed in Figure 4, in which the intervals do not overlap. In this case, we can rank the  $x$ -values (even though they're not precise) and assign them transformed grades based on their ranks. The regression plot for a normal distribution fitted to the skinny data is shown on the left graph of Figure 22. It seems less obvious, however, how the approach would be applied if the data intervals overlap, because the intervals cannot be sorted by magnitude. For instance, the intervals of the puffy data set displayed in Figure 5 cannot be uniquely sorted. We can nevertheless determine bounds on the smallest (scalar) value of this data set. What is it? It could certainly be any value in the interval  $x_7 = [0.15, 3.8]$ . But it might also be the value represented by  $x_4 = [2.8, 6.7]$ , or the one represented by  $x_5 = [3.5, 9.7]$ , or even the one represented by  $x_1 = [3.5, 6.4]$ . But, in any case, the smallest scalar value in the data set has to be a number no larger than 3.8, because  $x_7$  cannot be any larger than this value. We can find bounds on the second smallest value in the data set in a similar way. This value could surely be no smaller than 2.8 (the left endpoint of  $x_4$ ). It might be this value if there one value is in  $x_7$  (less than or equal to 2.8) and one value at the left edge of  $x_4$ . The largest magnitude the second smallest value in the data set could possibly have is 4.9, which is the right endpoint of  $x_8$ . Similar deductions can be used to bound the third smallest value in the data set, and so forth. These bounds on the ranked data seem complicated, but they are actually very easy to compute in practice. All we need to do is separately sort the left and right endpoints of the input data. The  $i^{\text{th}}$  value in the sorted list of left endpoints is the lower bound on the  $i^{\text{th}}$  smallest (scalar) value in the data set. An interval regression can then be done on these bounds on the  $i^{\text{th}}$  smallest value against the grades  $i/(N+1)$  transformed according to the shape of the distribution to be fitted. The regression plot for a normal distribution fitted to the puffy data set is shown on the right graph of Figure 22.

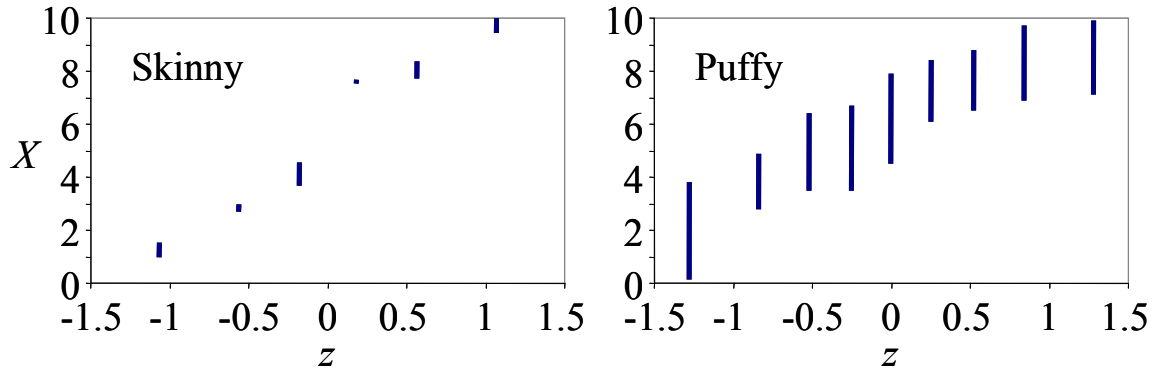


Figure 22. Regressions of  $X$  against normal scores.

### 4.9.3 Maximum likelihood methods

The idea behind the maximum likelihood approach to fitting a statistical distribution to a data set is to find the parameters of the distribution that maximize the likelihood of having observed the data. Assuming the data are independent of each other, the likelihood of the data is the product of the likelihoods of each datum. The likelihood of each datum is simply the probability associated with that value by the distribution if the parameters had been the actual ones.

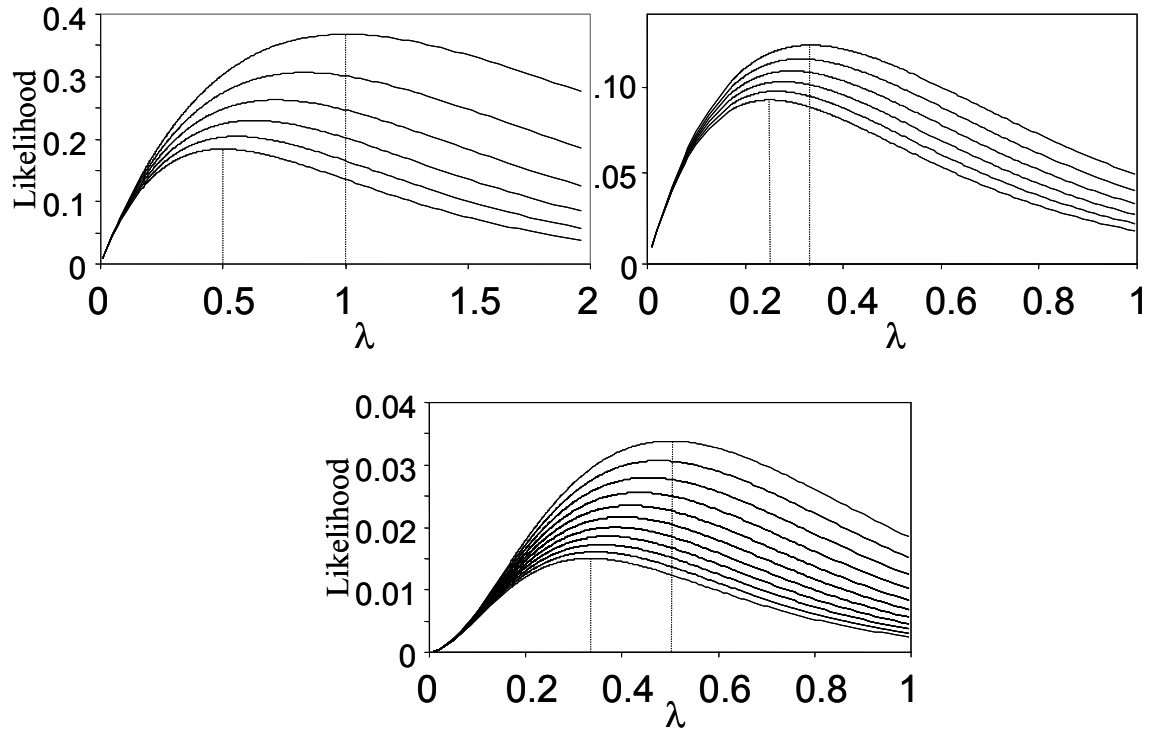
When the data are intervals rather than point estimates, there is no longer a single maximum likelihood distribution, but a class of them. For instance, suppose we have one data sample which is the interval  $x = [1, 2]$  and we want to find the maximum likelihood exponential distribution given this datum. The probability density function for an exponential distribution is

$$\lambda \exp(-\lambda x)$$

where  $1/\lambda$  is its parametric mean. We evaluate this quantity as a function of  $\lambda$ , and find a class of functions, a few of which are depicted in the upper, left-hand graph of Figure 23. There are many possible likelihood functions because we are unsure about the precise value of the datum represented by  $x$ . The graph reveals that the values of  $\lambda$  which have the highest likelihood range from  $1/2$  to  $1$ . Thus, the class of exponential distributions with  $\lambda \in [1/2, 1]$  would be the result of fitting an exponential distribution to the interval datum  $x = [1, 2]$  by maximum likelihood.

If there were instead two independent data points,  $x_1 = [1, 2]$  and  $x_2 = [3, 4]$ , the likelihood classes would be computed for each data interval separately. The two graphs at the top of Figure 23 show the range of resulting likelihood functions that would be obtained for each of these two data. The graph on the left shows the likelihood functions associated with the datum  $x_1$ , and the graph on the right shows the likelihood functions associated with the datum  $x_2$ . The likelihood functions within these two classes are pairwise multiplied (because the data are independent). The resulting Cartesian space of likelihood functions is depicted as the lower graph in Figure 23. The maximum

likelihood values of  $\lambda$  are in the interval  $[1/3, 1/2]$ . Thus, the class of exponential distributions with  $\lambda \in [1/3, 1/2]$  is the result of fitting an exponential distribution to the two data by the method of maximum likelihood to the two data intervals.



**Figure 23. Fitting an exponential distribution to two intervals by maximum likelihood.**

In this case, we could have simply used the formula for the maximum likelihood estimate for the exponential parameter  $\lambda$ , which is

$$\hat{\lambda} = \frac{1}{\sum_{i=1}^N x_i}$$

by plugging in the interval data. Because there are no repeated parameters in this expression, it yields the exact bounds on  $\lambda$ , which are  $([1,2]+[3,4])/2)^{-1} = [1/3, 1/2]$ , as previously seen.

But the formulas for the maximum likelihood estimates for parameters of other distributions are not always so simple. The formulas for the normal distribution specify the arithmetic mean and the (population) standard deviation, which, as we saw in section 4.9.1, are dependent on each other. As a result, as we saw with the method of matching moments, it is relatively easy to obtain conservative results by using the interval analogs for the arithmetic mean and standard deviation to obtain a bounding p-box. Further research is needed to demonstrate how this p-box can be tightened by taking account of the dependence between the mean and the standard deviation.

Likewise, the formulas for the maximum likelihood estimates for the parameters of a Weibull distribution, for instance, are the solutions to the simultaneous equations

$$\hat{b} = \left[ \frac{1}{N} \sum_{i=1}^n x_i^{\hat{c}} \right]^{1/\hat{c}}$$

$$\hat{c} = \frac{N}{\left(1/\hat{b}\right) \sum_{i=1}^N x_i^{\hat{c}} \log x_i - \sum_{i=1}^N \log x_i}$$

where  $N$  is the number of samples. Because  $\hat{b}$  and  $\hat{c}$  occur on both sides of the equal signs, they are not easy to solve for, even when the data are scalar values. When the input data  $x_i$  are intervals, the problem of solving for these maximum likelihood estimates becomes enormously difficult and, because the uncertain inputs  $x_i$  are repeated within and across these equations, their uncertainties are inserted into the evaluations many times leading to parameters with inflated uncertainties. Consequently, it does not seem feasible to apply these formulas to interval data sets. The problem will probably require that new formulas appropriate for interval data sets be derived. It is not necessarily the case, however, that the resulting formulas will be significantly more complex than those already used for scalar data.

*Numerical examples.* In the case of the exponential distribution, the maximum likelihood approach and the method of matching moments (section 4.9.1) yield the same result. Thus, the results of fitting the skinny and puffy data sets to an exponential distribution via maximum likelihood are displayed in Figure 19.

Meeker and Escobar (1995) describe an alternative maximum likelihood approach commonly employed for interval-censored data that is not formulated in terms of a class of possible likelihood functions. Instead, it computes a single likelihood in terms of the probability of *being within an interval*. This approach is widely used and is important because it provides analysts with a convenient way to handle interval data. However, it is not compatible with the approach to incertitude adopted in this report. To see why we might object to the approach, consider the following simple, albeit rarefied, example of fitting a normal distribution with unit variance to the interval datum  $[\underline{x}, \bar{x}] = [1, 2]$ . The likelihood of the normal distribution with parameters  $\mu$  and  $\sigma=1$  would be

$$\Phi\left(\frac{\bar{x}-\mu}{\sigma}\right) - \Phi\left(\frac{\underline{x}-\mu}{\sigma}\right) = \Phi(2-\mu) - \Phi(1-\mu)$$

where  $\Phi$  is the standard normal cumulative distribution function. This difference of the cumulative probabilities gives the probability, under the model, of being in the interval  $[1, 2]$ . This approach produces a unique solution to the maximum likelihood problem, which is the distribution centered on the interval and defined by  $\mu = (\underline{x} + \bar{x})/2 = (2 + 1)/2 = 1.5$ .

Although proponents of this approach argue that it accounts for the interval uncertainty and does not make any distributional assumption within the interval, we see these advantages as insufficient for the purpose of propagating uncertainty. The fact that it produces a *single* result strongly violates our intuition about the uncertainty represented by the interval. We argue that this approach creates a unique\* solution where none is justified. Presuming that the interval represents uncertainty as characterized in this report, there is surely nothing to require the normal distribution centered at the interval midpoint to be any more likely, given the evidence, than one centered at any other point within the interval. Because there is nothing to distinguish the possibility that the datum's true value—if we could measure it precisely—might be 1.5 from the possibility that is 1.25, or 1.001, or any value between 1 and 2, there should be nothing to distinguish the centered distribution as having more likelihood in any sense that is informed by the evidence than any distribution with a mean at another point in the interval. The assumption of normality cannot, by itself, create this result. There is clearly some symmetry or uniformity principle in play here, even if it is not immediately recognizable. Defenders of this approach might suggest that it is okay for the estimate to understate the uncertainty about the best fit distribution because the result is usually expressed along with confidence bounds of some kind. This is the same defense that could be mounted for any of the statistical methods that neglect uncertainty, and it seems unconvincing to us. Once the uncertainty has been lost through the creation of a precise likelihood function, it cannot really be recovered in subsequent analysis or disquisition based solely on that likelihood function.

Although this alternative approach to interval-censored data has been useful in introducing interval data to statisticians and giving them a strategy to deal with such data, it is not compatible with the perspectives adopted in this report. In short, it *handles* interval uncertainty but does not *capture and propagate* it in a way doesn't add additional structure to the problem beyond what is embodied in the intervals and the admitted assumptions (in the case of the example above, the normality assumption). Fortunately, as we saw in the first part of this section, other maximum likelihood approaches can be constructed that fully express and account for uncertainty in data.

To respect the uncertainty in interval data, we want to be conscientious in not inadvertently introducing assumptions about where a value might lie within its interval, but this does not preclude our drawing inferences a posteriori about the probability distribution for the value within an interval. For instance, suppose that the seventh interval in the puffy data set (page 28) were replaced by a much wider interval [0.15, 100] which would lead to a mean for the data set of [4.56, 18.08]. These bounds on the mean have implications for the probability of the datum within the wide seventh interval. We know, for instance, that the chance that a value drawn at random from an exponential distribution, even with a mean as big as 18.08, being larger than 50 is less than 7%.

---

\*If the width of the interval is large compared to the dispersion  $\sigma$ , the likelihood function may have a flat top so that there is a range of means that share the same maximal likelihood of unity, but this range is always strictly smaller than the interval itself, which is contrary to the intuition about what is likely given the evidence. In such cases, the likelihood of  $\mu$  at the endpoints of the interval is always one half the likelihood associated with  $\mu$  at the interval's midpoint.

Intuitively, therefore, we can say something about the probability distribution within this wide seventh interval after considering the data, even though we could not say anything when it was first recorded. The proper inference about such chances is subtler than this because it must also account for sampling uncertainty and our having only 9 data values from which the mean is estimated. Although a full discussion of this topic is beyond the scope of this report, it is clear that the other data, once empowered by the assumptions of randomness and the shape of the distribution, can tell us something about what's going on inside the wide interval. Likewise, mutual inferences about each interval in the data set relative to the other data are also possible. Of course, these inferences depend crucially on the data being truly randomly drawn from an exponential distribution, and they do not come for free through the use of an overly precise formulation of the likelihood function.

#### 4.9.4 Other methods for fitting distributions

There are other nonparametric methods for fitting distributions to data. These methods are nonparametric in the sense that they do not require the analyst to *name* a particular shape for the distribution. We mention two of these methods.

The most popular nonparametric method for estimating probability distributions is kernel smoothing (Ward and Jones 1995). Given a set of scalar data  $X_i, i = 1, \dots, n$ , from a continuous univariate probability density function  $f$ , the kernel estimator of the density function is

$$\hat{f}(x, h) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right)$$

where  $h$  is the “bandwidth” that controls the smoothness of the estimate and  $K$  is the kernel function. The kernel can be a Gaussian density function, the uniform density function, or some other continuous, bounded, symmetric real function centered at each datum  $X_i$ . Because kernel smoothing can show arbitrarily high fidelity to the available data depending on the value of the bandwidth used, the method can uncover structure in a data set that parametric methods might miss. But, unlike the empirical distribution function of the data, the result of kernel smoothing can have tails that extend beyond the range of the observed data. Because the kernel smoothing estimator is essentially a mixture (vertical average) of  $n$  kernel functions and the mixture operation extends in a straightforward way when the input data contain interval uncertainty, this method can easily be generalized to the case of interval data.

Beer (2006) describes a novel method of iterated constrained bootstrapping with random modification by which a small sample (~200 values) may be used to generate an arbitrarily large sample, which is essentially equivalent to defining a distribution function, based on the original sample. He argues that the method has properties that suggest the method is preferable to ordinary bootstrapping and kernel smoothing methods. Beer shows that this method can be extended immediately to interval data sets.

## 5 Inferential statistics with interval data

Inferential statistics refers to those methods by which analysts draw conclusions about statistical populations from samples observed from those populations. Inference is distinguished from mere quantitative description of the observations (which was covered in section 4). Confidence limits, which are considered by some to be within the realm of inferential statistics, were already addressed in sections 4.1.2 and 4.8. This section introduces the inference problems of testing the equality of variances, detecting outliers, and linear regression for data sets containing uncertainty. For the last of these topics, the discussion is only an overview and does not offer solutions to the computational problems.

Statistical inference using hypothesis testing is widely used in engineering. Statistical methods are needed because the underlying processes that generate the patterns in data are often multifactor, complicated, and usually unknown. Virtually no serious scientist or engineer today believes Lord Rutherford's quip that, if you need statistics, you ought to have done a better experiment. The questions we are asking today are harder, and we need to wring from every hard won datum whatever information it has. Statistical inference gives analysts a way to demonstrate that something other than random chance is at work in making the patterns seen in data from an experiment.

Classical Neyman-Pearson hypothesis testing uses traditional scientific negation in positing a hypothesis and garnering evidence to reject it if possible. A hypothesis test is a procedure for deciding between alternate hypotheses on the basis of the value of a test statistic which is a function of observations in some random sample. Usually, one of the hypotheses holds that there is no pattern in the data beyond what is there by random chance. This is called the null hypothesis, and it corresponds to the possibility that there is no effect that would be measured by the test statistic. If the test statistic is far from the values it might have by random chance under the null hypothesis, the test rejects the null hypothesis. If the test statistic has a value close to those values, the test accepts, or rather fails to reject, the null hypothesis. A type I error occurs when the null hypothesis is true but the test rejects it. The type I error rate is conventionally controlled at the threshold probability level  $\alpha = 0.05$ . If the probability computed for a statistical test is below this level, then the null hypothesis of no effect is rejected and the effect is said to be statistically significant. Even if there actually is an effect present, the rejection of the null hypothesis is not guaranteed if there is large sampling uncertainty or measurement uncertainty or both. If, however, the probability is above the threshold level, the null hypothesis fails to be rejected. The reason for this outcome might be either because the null hypothesis (that there is no effect) is true, or because the evidence against it was so slim that it did not look any stronger than fluctuations that might have arisen by random chance alone.

The conventional statistical strategy for assessing measurement uncertainty (mensural imprecision) in statistical tests is to take repeated measurements and treat the imprecision as an additional source of variation. But the labor-intensive nature of experimentation in



the laboratory and especially in the field often limits the ability of analysts to collect sufficiently many samples sizes to achieve an acceptable level of statistical power. Consequently, repeated samples are often sacrificed first, and many if not most experimental designs do not take measurement uncertainty into account. For the many experiments in which there are no repeated measurements, the effect of measurement imprecision is wholly unexplored. For such experiments, one cannot tell whether measurement uncertainty can have an effect on the results of hypothesis testing.

Using intervals to express and propagate the mensural uncertainty associated with each datum is a relatively inexpensive way to correct this situation and make a formal accounting of the impact of measurement uncertainty. In many cases, this can be done in an entirely straightforward way by putting intervals around measurements, and using the methods of interval analysis to compute bounds on the test probability. When the resulting bounds are to one side or the other of the threshold  $\alpha$  level, the decision of the hypothesis test is clear. Insofar as it accounts for uncertainties that would not have otherwise been accounted for, an approach using interval statistics increases the credibility of the decision. When the bounds on the test probability straddle the threshold  $\alpha$  level, the implication is that the test is inconclusive because of the mensural imprecision in the data. Sometimes this result can be remedied by collecting more data samples, but sometimes it cannot be, simply because getting more poor information (data with wide intervals) may not lead to more surety about the result. In such cases, it may be necessary to improve the measurements themselves rather than merely getting more of them.

## 5.1 Equality of variances for imprecise data

A statistical test of the null hypothesis that the variances are equal in two groups of normally distributed random samples can be made with the test statistic

$$\frac{s_2^2}{s_1^2},$$

where  $s_1^2$  and  $s_2^2$  are the respective sample variances, labeled so that  $s_1^2 < s_2^2$ . The statistic is compared against critical values of the  $F$  distribution (also called the Fisher-Snedecor or variance-ratio distribution) with degrees of freedom that depend on the sample sizes in the two groups. These critical values are tabled in most statistics textbooks (e.g., Crow et al. 1960, Table 5, page 234ff). When the test is extended to account for interval data, the variance estimates and the ratio test statistic are intervals as well.

*Numerical example.* We saw in section 4.5.1 that the sample variances for the skinny and puffy data sets are [9.503, 12.912] and [1.03, 12.35] respectively. The test for equality of variance computes its statistic as their ratio [9.503, 12.912] / [1.03, 12.35] = [ 0.76, 12.54], with outward directed rounding. This interval can be compared to the critical value of the  $F$  distribution with degrees of freedom for the denominator  $9 - 1 = 8$ , and

degrees of freedom for the numerator  $6 - 1 = 5$ , which is the value 4.818, or 6.757 for a two-tailed test. In this case, the uncertainty in the data sets prevents us from concluding whether or not the difference between the variances of the two data sets is statistically significant at the 0.05 level.

*Another numerical example.* Suppose there are two data sets with 56 and 96 samples each, for which the interval variances are  $s_1^2 \in [100.7, 116.3] \text{ kg}^2$  and  $s_2^2 \in [231.5, 257.8] \text{ kg}^2$ , respectively. These variance intervals constrain the resulting ratio statistic to the interval  $[1.990, 2.561]$  with outward-directed rounding. The statistic should be compared with tabled critical values of the  $F$  distribution with degrees of freedom  $df_1 = 55$  and  $df_2 = 95$ . Because the tabled critical value of the  $F$  distribution is a monotonically decreasing function of the degrees of freedom (so long as  $df_2 > 2$ ), we can infer that the bounds on the type I error rate are  $p \in [0.00067, 0.018]$ . If the test is two-tailed, as it usually is, the probability is doubled so  $p \in [0.00134, 0.036]$ . In this case, the null hypothesis that the masses in the two populations are homoscedastic (have equal variances) is rejected.

## 5.2 Detecting outliers

Outlier detection has numerous applications in science and engineering. In risk assessment, statistical outliers may constitute an early warning associated with extreme events. In materials testing, outliers may be indications of possible faults and future failures. Detecting outliers is also an important function in bioinformatics (Shmulevich and Zhang 2002). Likewise, the search for mineral and oil deposits is often conducted as a statistical search for outliers in survey data. The importance of interval uncertainty for such applications was emphasized by Nivlet et al. (2001a,b).

The implication of a value's being considered an outlier is that the value is so unusual, so departed from the norm, that it must have been produced by an entirely different process than that which created other values also observed. Recognizing a value as an outlier requires characterizing the statistical distribution from which it was supposedly drawn. The problem of characterizing distributions was addressed in sections 4.1 and 4.9. Of course, the information available to characterize the distribution includes the would-be outlier, so the considerations become a bit circular. Conventionally in engineering, a value is called an outlier if it is greater than some number of standard deviations from the mean of the distribution. Thus, the inferential statistics problem of outlier detection is closely related to the descriptive statistics problem of calculating confidence intervals (see section 4.8).

For every datum  $x_i$ , we can determine the degree of 'outlierness'  $r$  as the smallest  $k_0$  for which  $x_i \notin [E - k_0\sigma, E + k_0\sigma]$ , where  $E$  is the expectation (arithmetic average) and  $\sigma$  is the standard deviation of the underlying distribution. Hence,  $r = |x_i - E|/\sigma$ , which is a commonly used ratio in statistics. We can simplify its calculation in two ways. First, the ratio is invariant to shift translations of the data, which is to say, it remains the same if we add or subtract any fixed number to all the  $x$  values, because both the difference in the numerator and the standard deviation in the denominator are unchanged when shifted by a constant. Without losing generality, we can assume that  $x_i = 0$ , and we are therefore

interested in the ratio  $|E|/\sigma$ . Second, the lower bound of the ratio  $r$  is attained when the inverse ratio  $1/r = \sigma/|E|$  is the largest, and vice versa. Thus, to find the interval of possible values for  $|E|/\sigma$ , it is necessary and sufficient to find the interval of possible values of  $\sigma/|E|$ . Computing this interval is, in its turn, equivalent to computing the interval for the square  $V/E^2$  of the reverse ratio  $1/r$ . Finally, since  $V = M - E^2$ , where  $M = (x_1^2 + \dots + x_N^2) / N$  is the second moment, we have  $V/E^2 = M/E^2 - 1$ , so computing the sharp bounds for  $V/E^2$  is equivalent to computing the sharp bounds for the ratio  $R = M/E^2$ .

If you are skipping the derivation of algorithms, you can resume reading with section 5.3 on page 94.

### 5.2.1.1 Lower bound on $R$

Computing the lower bound  $\underline{R}$  can be accomplished with the following idea. For every  $i$ , the location of the minimum on the interval  $[\underline{x}_i, \bar{x}_i]$  depends on the values of the derivative

$$\frac{\partial R}{\partial x_i} = \frac{2}{N \cdot E^2} \cdot \left( x_i - \frac{M}{E} \right).$$

Thus, once we know where  $\lambda = M/E$  is located in comparison with the endpoints, we can uniquely determine all the values of the extremal configuration  $X_i$ —and the value  $\lambda$  can be determined from the condition that the ratio  $M/E$  is exactly equal to  $\lambda$ . Thus, we arrive at the following algorithm.

*Algorithm for computing lower bound on outlieriness ratio  $R$ .*

1. Data are  $x_i = [\underline{x}_i, \bar{x}_i]$ ,  $i = 1, \dots, N$ , which have no common intersection,  $\bigcup x_i = \emptyset$ , (otherwise, the variance can be zero and the lower bound  $\underline{R} \equiv 1$ ).
2. Sort all  $2N$  values  $\underline{x}_i, \bar{x}_i$  together into a sequence  $Y_1 \leq Y_2 \leq \dots \leq Y_{2N}$ , and assign  $Y_0 = -\infty$  and  $Y_{2N+1} = +\infty$ . These values divide the real line into  $2N+1$  zones  $(Y_0, Y_1]$ ,  $[Y_1, Y_2]$ ,  $\dots$ ,  $[Y_{2N-1}, Y_{2N}]$ ,  $[Y_{2N}, Y_{2N+1})$ .
3. For each of these zones  $[Y_k, Y_{k+1}]$ ,  $k = 0, 1, \dots, 2N$ , compute the values

$$\begin{aligned} N_k &= \# \{i : \bar{x}_i \leq Y_k\} + \# \{j : \underline{x}_j \geq Y_{k+1}\}, \\ e_k &= \sum_{i: \bar{x}_i \leq Y_k} \bar{x}_i + \sum_{j: \underline{x}_j \geq Y_{k+1}} \underline{x}_j, \\ m_k &= \sum_{i: \bar{x}_i \leq Y_k} \bar{x}_i^2 + \sum_{j: \underline{x}_j \geq Y_{k+1}} \underline{x}_j^2. \end{aligned}$$

4. Compute  $\lambda_k = m_k / e_k$ .
5. If  $\lambda_k \in [Y_k, Y_{k+1}]$ , then compute

$$E_k = \frac{e_k}{N} + \lambda_k \frac{N - N_k}{N} \quad M_k = \frac{m_k}{N} + \lambda_k^2 \frac{N - N_k}{N}$$

and  $R_k = M_k / E_k^2$ .

6. The smallest of these  $R_k$  values is the desired lower bound on  $R$ .

Sorting requires  $\mathcal{O}(N \log N)$  time, and computing the initial values of  $e_k$ ,  $m_k$ , and  $N_k$  requires  $\mathcal{O}(N)$  time. Updating each of these values requires a constant number of steps, so the overall time is linear. Thus,  $\underline{R}$  can be computed in  $\mathcal{O}(N \log N)$  time.

*Numerical example.* Just to make sure the mechanics are clear, let us trace through the algorithm with an example with a sample size of two, and data  $x_1 = [1,2]$ , and  $x_2 = [3,4]$ . In this case,  $Y_1 = 1$ ,  $Y_2 = 3$ ,  $Y_3 = 3$ ,  $Y_4 = 4$  and there are five zones:  $(-\infty, 1]$ ,  $[1,2]$ ,  $[2,3]$ ,  $[3,4]$ , and  $[4, +\infty)$ . Applying the algorithm for each zone:

- For  $(-\infty, 1]$ , we have  $e_0 = 1+3 = 4$ ,  $m_0 = 1^2 + 3^2 = 10$ ,  $N_0 = 2$ , hence  $\lambda_0 = 10/4 = 2.5$  which is outside its zone.
- For  $[1,2]$ ,  $e_1 = 3$ ,  $m_1 = 3^2 = 9$ ,  $N_1 = 1$ , hence  $\lambda_1 = 9/3 = 3$  which is outside its zone.
- For  $[2,3]$ ,  $e_2 = 3+2 = 5$ ,  $m_2 = 3^2 + 2^2 = 13$ ,  $N_2 = 2$ , so  $\lambda_2 = 13/5 = 2.6$  which is within the zone. In this case,  $E_2 = 5/2$ ,  $M_2 = 13/2$ . Hence

$$R_2 = \frac{13/2}{(5/2)^2} = \frac{13 \cdot 4}{25 \cdot 2} = \frac{52}{50} = 1.04.$$

- For  $[3,4]$ ,  $e_3 = 2$ ,  $m_3 = 2^2 = 4$ ,  $N_3 = 1$ , so  $\lambda_3 = 4/2 = 2$  which is outside the zone.
- For  $[4, \infty)$ ,  $e_4 = 2+4 = 6$ ,  $m_4 = 2^2 + 4^2 = 20$ ,  $N_4 = 4$ , so  $\lambda_4 = 20/6 = 3.33$  is outside the zone.

Because the only value  $R_k$  is 1.04, the lower bound  $\underline{R} = 1.04$ .

### 5.2.1.2 Upper bound on $R$ in the general case

In principle, the upper bound of  $R$  could be infinite if  $0 \in [\underline{E}, \bar{E}] = [\Sigma \underline{x}_i, \Sigma \bar{x}_i]$ . If the interval for  $E$  does not contain 0, for instance, if  $0 < \underline{E}$ , then  $\bar{R}$  is guaranteed to be finite, and we can bound it by the ratio  $\bar{M} / \underline{E}^2$ . When  $\bar{R} < N$ , the maximum  $\bar{R}$  is always attained at the endpoints at a corner of the Cartesian product of the data intervals (Kreinovich et al. 2005a). (This is a reasonable presumption because sample size is almost always more than a handful, and values of  $R$  that are any bigger than five or six would be very strong evidence for an outlier.) We can find the bound by testing all  $2^N$  combinations of  $\underline{x}_i$  and  $\bar{x}_i$ . This is an NP-hard algorithm.

### 5.2.1.3 Upper bound on $R$ for data with few intersections

This section gives a more practical algorithm due to Kreinovich et al. (2005a) to compute the upper bound on  $R$  for interval data in which there are fewer than  $K$  intervals that overlap at any point. It also works for interval data without any intersections of course, and in fact it works in a slightly more general case in which particularly narrowed intervals don't intersect, or no more than  $K$  have a common intersection.

*Algorithm for computing upper bound on outlieriness ratio  $R$ , assuming  $R < N$ .*

0. Data are  $x_i = [\underline{x}_i, \bar{x}_i]$ ,  $i = 1, \dots, N$ , such that  $0 \notin [E, \bar{E}] = [\Sigma \underline{x}_i, \Sigma \bar{x}_i]$  and no more than  $K$  intervals  $[\underline{x}_i^-, \bar{x}_i^+]$  have a common intersection, where

$$x_i^- = \frac{\underline{x}_i + \bar{x}_i}{\left(2 + \frac{\bar{x}_i - \underline{x}_i}{NE}\right)}, \text{ and } x_i^+ = \frac{\underline{x}_i + \bar{x}_i}{\left(2 - \frac{\bar{x}_i - \underline{x}_i}{NE}\right)}.$$

1. First sort all  $2N$  values  $\underline{x}_i, \bar{x}_i$  together into a sequence  $Y_1 \leq Y_2 \leq \dots \leq Y_{2N}$ , and assign  $Y_0 = -\infty$  and  $Y_{2N+1} = +\infty$  to divide the real line into  $2N+1$  zones  $(Y_0, Y_1], [Y_1, Y_2], \dots, [Y_{2N-1}, Y_{2N}], [Y_{2N}, Y_{2N+1})$ .
2. For each of these zones  $[Y_k, Y_{k+1}]$ ,  $k = 0, 1, \dots, 2N$ , and for each variable  $i$ , take
  - a.  $X_i = \underline{x}_i$  if  $x_i^+ \leq Y_k$ ,
  - b.  $X_i = \bar{x}_i$ , if  $x_i^- \geq Y_{k+1}$ , or
  - c. Otherwise both values  $\underline{x}_i$  and  $\bar{x}_i$ .
3. For each of these combinations, compute  $E, M$ , and  $\lambda = M/E$ , and check that  $\lambda$  is within its zone. If it is, then compute  $R_k = M/E^2$ .
4. The largest  $R_k$  so computed is the desired upper bound on  $R$ .

If this algorithm is implemented along the lines of the analogous algorithm for the upper bound on variance, it can also perform in  $\mathcal{O}(N \log N)$  time.

The algorithm is applicable when the ‘narrowed’ intervals don’t intersect or don’t intersect too much, but an intuitive understanding of how much overlap is possible is hard to develop. Consider a *family of data sets* that vary in their overall incertitude from perfect precision of point values to vacuously overlapping intervals. Figure 24 depicts such a family of data sets as six sideways triangles corresponding to  $N=6$  intervals in the data set. The points of the six triangles on the far left of the graph represent the perfect point data, and any vertical slice through these six triangles corresponds to an interval data set  $x_1, \dots, x_6$  of some overall precision from the family. For the leftmost 20% of the graph, the data intervals  $x_1, \dots, x_6$  don’t intersect at all, so they could certainly be handled by the algorithm. But, beyond this observation, it is impossible to discern by inspection how much overlap among the data intervals the algorithm can tolerate before it breaks down. For this example, the unshaded portion of the triangles (the leftmost 60% of the graph) represent data sets whose intervals narrowed in the sense of the algorithm do not intersect. The region with light gray shading corresponds to data sets in which the narrowed intervals intersect but only by pairs, so the algorithm applies with  $K = 3$ . Combining the unshaded and light-gray-shaded regions covers about 95% of the data sets depicted in the graph. Beyond this degree of incertitude, the region of the graph with dark gray shading represents data sets in which triple-wise intersections of the narrowed intervals start to appear. The point of the figure is that interval data can have a

considerable amount of overlap and still be amenable to analysis by the algorithm, although the degree of overlap that can be tolerated is influenced by how evenly scattered or clustered the midpoints of the intervals happen to be.

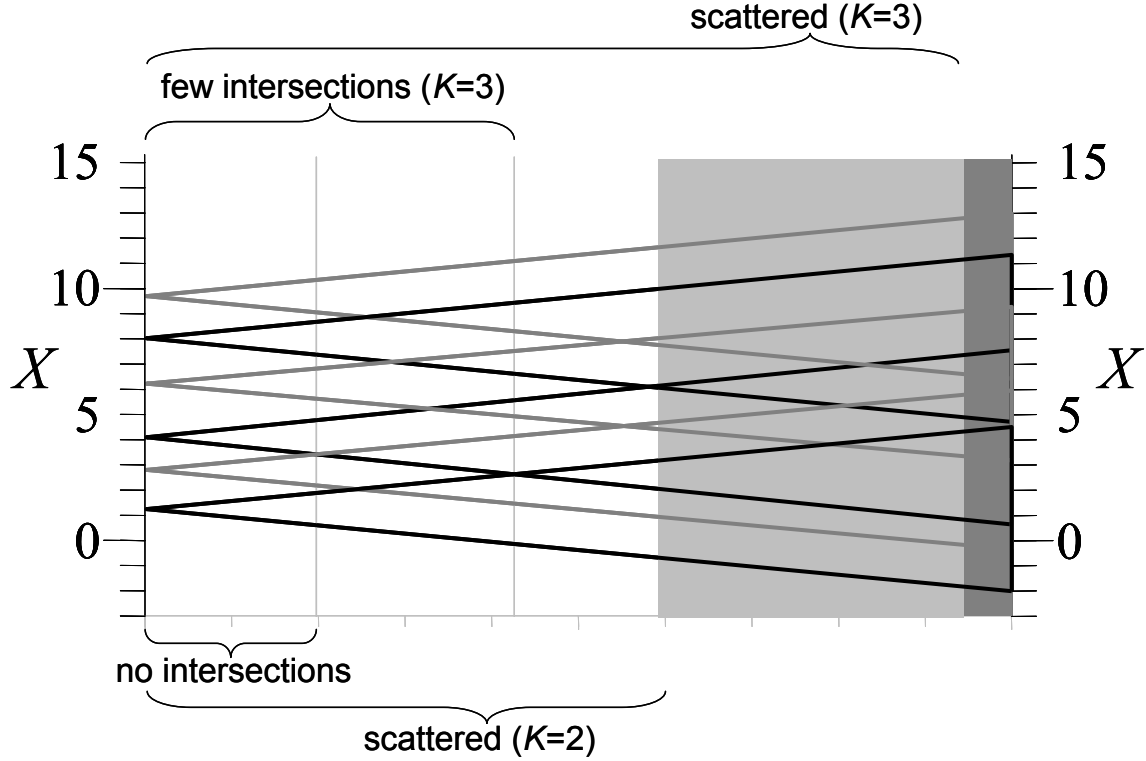


Figure 24. Uncertainty (vertical widths) growing from left to right in a data set ( $N=6$ ) until ‘narrowed’ intervals intersect (light gray) and triple-wise intersect (dark gray).

#### 5.2.1.4 Upper bound on $R$ for other special data cases

For no-nesting interval data such as we would expect when they are obtained from a single measurement device, the problem of computing the upper bound on  $R$  is similar to the problem of computing the upper bound on the variance (see section 4.5.1.2). We can prove that the maximum  $R$  is attained on one of the vectors  $x = (\underline{x}_1, \dots, \underline{x}_K, \bar{x}_{K+1}, \bar{x}_N)$  for some  $K$  in some ordering of the indices. It is thus possible to compute the upper bound in  $\mathcal{O}(N \log N)$  time. Likewise, for the case several-instruments interval data from  $M$  measuring devices, we can compute the upper bound on  $R$  in  $\mathcal{O}(N^m)$  time. Computing the upper bound on  $R$  for other data cases is similar in complexity to computing bounds for the confidence limits  $L$  and  $U$  (see section 4.8), and the computational problems have similar complexity estimates.

### 5.3 Linear regression

Interval regression has already been mentioned in section 4.9.2 in the context of fitting distributions to data. This section will not present a complete discussion of the topic of

interval regression nor even outline the algorithms that might compute an interval regression, but only review a few of the most salient issues that it involves. Several researchers have described the application of linear regression to data sets that contain interval uncertainty (Markov 1990; Manski and Tamer 2002; Marino and Palumbo 2002; 2003). Generalizations of regression for fuzzy data have also been described (Klir and Yuan 1995; Salla and Ferson 1998), and, because fuzzy data are generalizations of interval data, these methods can be applied immediately to interval data. Principal components analysis, sometimes called “model II regression”, which differs from the traditional least squares method in its error model, has also been generalized for data containing interval uncertainty (Lauro and Palumbo 2000; 2003; 2005; Gioia and Lauro 2006; Lauro et al. 2007).

There are different ways in which regression can be generalized to account for incertitude in data. One way is to compute intervals for the  $p$ -values that could be obtained from ordinary least squares regression when the input values are allowed to freely vary within their prescribed interval ranges. If the  $p$ -values are all below the 0.01 level, for instance, one might conclude that the data clearly exhibit a significant regression despite the magnitude of the incertitude that they may contain. Likewise, if the  $p$ -values are all above 0.05, then one would conclude there is little evidence for a statistical signal in the data. On the other hand, if the  $p$ -values straddle a critical level, one might reasonably conclude that the uncertainty in the data prevents our making a clear determination about the regression.

A regression analysis generalized for interval data can also compute intervals for the slope and intercept from the collection of all regression lines, each of which is best fit to some configuration of scalar points. It may be important to keep in mind that these slopes are going to be dependent on the intercepts. If this dependence is important to the analyst, it will be reasonable to display the incertitude about the slope and intercept with a two-dimensional display rather than separate intervals. The dependence will be manifested as the possible combinations of slope and intercept filling only a proper subset of the rectangle implied by the respective intervals.

It may sometimes be useful to depict the envelope of possible regression lines, each one of which is the result of an ordinary linear regression analysis based on some configuration of bivariate scalar points from within the intervals. Figure 25 illustrates the idea. In this case, the envelope was estimated with a Monte Carlo optimization strategy involving 100,000 iterations, so the result is not guaranteed in the sense of a rigorous interval or constraint analysis, but is probably pretty good.

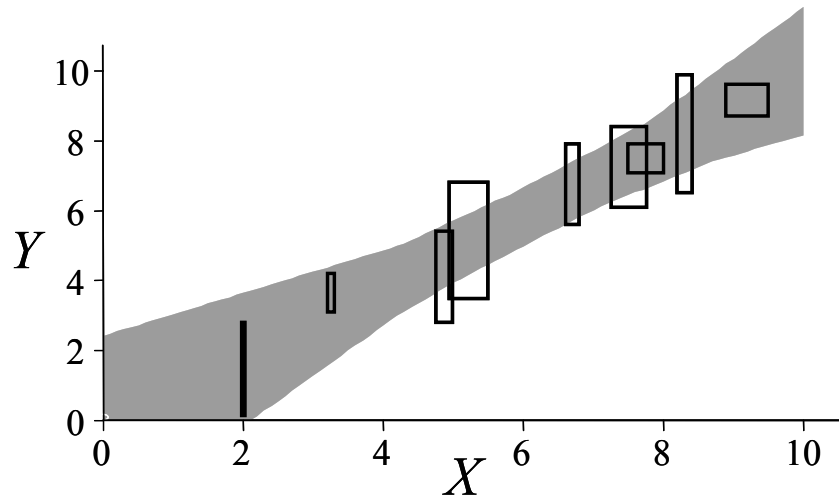


Figure 25. Envelope of regression lines for interval data.

Regressions with interval data also allow a kind of analysis that has no analog among traditional statistical methods. Consider the collection of lines (or other functional shapes, for that matter) that pass through *every* datum. These lines are not statistical in the traditional sense, but rather they are the result of a constraint analysis. An *all-constraints* regression treats every interval datum as a constraint and finds the set of lines that satisfies all of them. Figure 26 shows an example of such an analysis with 9 bivariate intervals. Only a few lines from the pencil (collection of lines) that pass through all nine regions are depicted, but they include the lines with the largest and smallest possible slopes and the largest and smallest possible intercepts. The collection of lines from an all-constraints regression will be non-empty whenever the largest correlation for the interval data is unity.

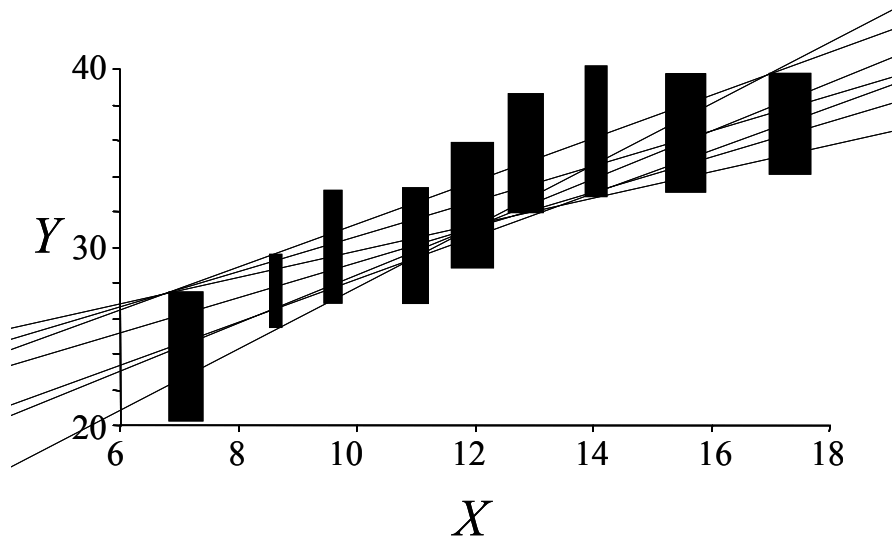
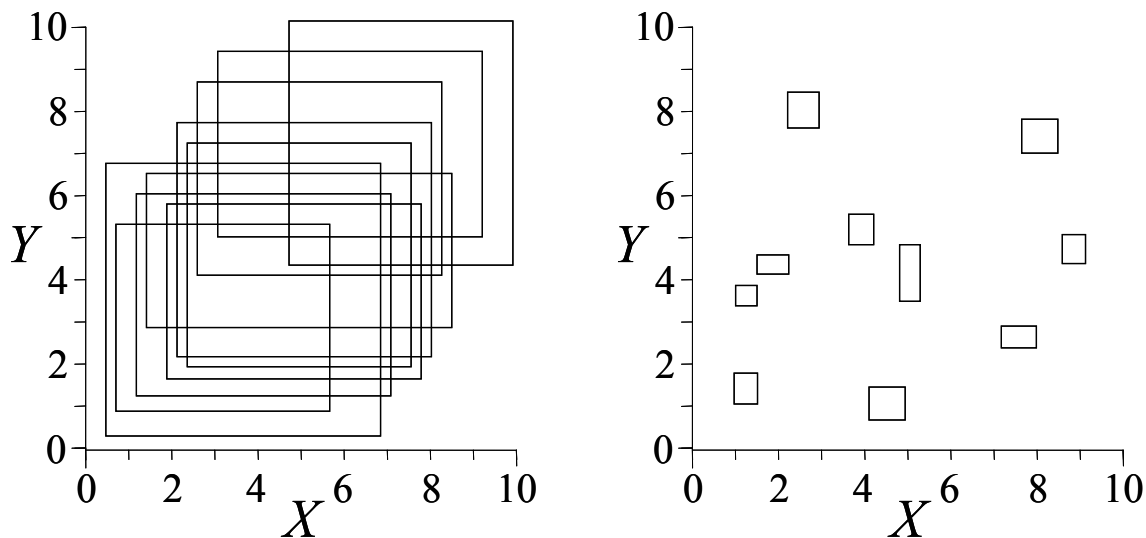


Figure 26. An example of an all-constraints regression.



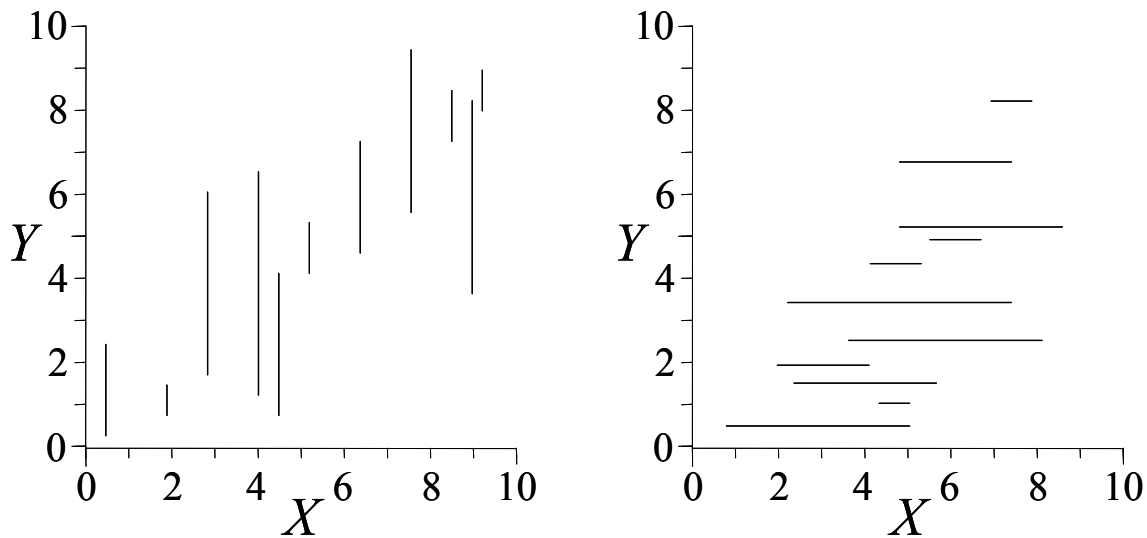
Figure 27 depicts two hypothetical bivariate data sets with interval uncertainty. These graphs illustrate two distinct reasons why a data set may fail to contain significant evidence for a linear regression. The spherical scatter of the data in the right graph lacks a trend. Such a scatterplot would lead to a regression line with a slope near zero, or near unity if model II regression is used. The graph on the left also lacks evidence for a linear regression, but the reason is not the absence of a discernible trend (of sorts), but rather the lack of precision in the input data. Notice that if the rectangles on the left graph were replaced by their centroids, such as would be done in a traditional statistical analysis, the resulting regression could easily be statistically significant because those centroids tend to align along a particular line. Yet that line would be rather misleading as a representation of the trend in the data given their acknowledged incertitude, which could, after all, readily accommodate a line with either positive or negative slope in literally any direction (because all ten bivariate intervals intersect). A traditional linear regression will detect a signal no matter how unreliable the data points are if they happen to line up. It is sensitive to only two features of the data: their collinearity, and their count. Indeed, if there are sufficiently many points in a data set, statistical significance of the regression analysis is virtually assured. In the past, when data were hard to come by, this was rarely a problem. But in the modern technological world where data collection is automated and conducted at high speed, it may become necessary to distinguish good data from poor data in the way that is now possible by representing data with intervals.



**Figure 27. Two reasons a regression may lack a strong signal.**

Figure 28 also depicts two hypothetical bivariate data sets with interval uncertainty. In these cases, the uncertainty is restricted to only one of the two variables. On the left graph, the uncertainty is in the  $Y$  variables while the  $X$  variables are known perfectly as scalars. On the right graph, the uncertain and precise variables are reversed. These two cases are clearly special cases of the general situation in which both variables have incertitude. It is useful to distinguish these special cases because they enjoy computational economies over the general case. It should be mentioned that ordinary linear regressions on scalar data follow a similar classification based on which of the two

variables is measured with uncertainty. In classical least squares regression, the independent variable is presumed to be measured without error and the dependent variable is presumed to have uncertainty (characterized as normal distribution). For this problem, the vertical distances between the points and the regression line are minimized. This structure is reversed in the so-called Berkson case (Sokal and Rohlf 1981), in which the  $Y$  variable is measured perfectly but the  $X$  variable has distributional uncertainty. In this problem, the *horizontal* distances between the points and the line are minimized. When both independent and dependent variables are subject to errors in measurement of roughly comparable magnitude, Model II regression (Sokal and Rohlf 1981) such as principal components analysis is used, which minimizes the *perpendicular* distances between the line and the points. In all of these models, however, any uncertainties in measurement are characterized by probability distributions. If the measurement uncertainty captured by interval data cannot really be represented by distributions, then the classical methods need the generalizations for interval data that have recently been described (inter alia, Klir and Yuan 1995; Saila and Ferson 1998; Lauro and Palumbo 2000; 2003; 2005; Manski and Tamer 2002; Marino and Palumbo 2002; 2003).

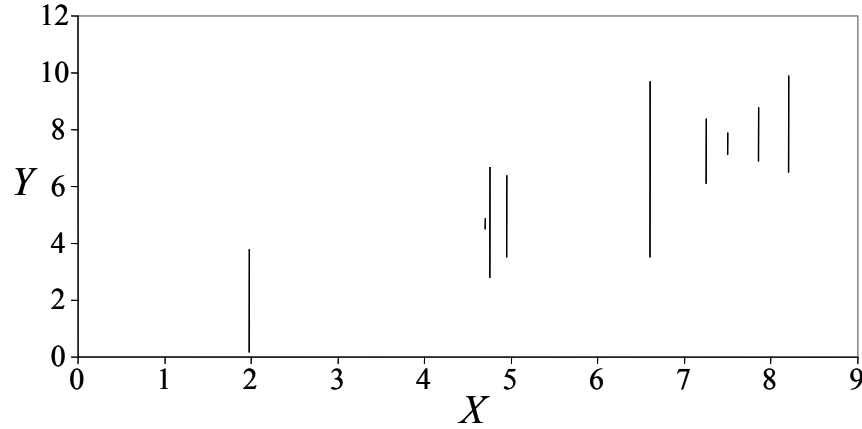


**Figure 28. Bivariate linear regressions with uncertainty in one only variable.**

*Numerical example.* Suppose we have collected the following paired precise  $X$  values and uncertain  $Y$  values:

$X$	$Y$
4.95	[3.5, 6.4]
7.85	[6.9, 8.8]
7.25	[6.1, 8.4]
4.75	[2.8, 6.7]
6.6	[3.5, 9.7]
8.2	[6.5, 9.9]
1.975	[0.15, 3.8]
4.7	[4.5, 4.9]
7.5	[7.1, 7.9]

This bivariate data set is displayed in Figure 29. The least squares slope ranges on  $[0.363, 1.637]$  for these interval data. The intercept of the regression is  $[-3.597, 3.598]$ . These intervals were computed with the optimization techniques in Microsoft Excel's Solver (Fylstra et al. 1998).



**Figure 29. Data for the numerical example.**

*Nonparametric regression.* There are also several nonparametric methods for regression analysis (Rousseeuw and Leroy 1987; Siegel 1982). One that is especially convenient to compute is the median of pairwise slopes

$$\frac{Y_j - Y_i}{X_j - X_i}$$

over all data pairs,  $1 \leq i < j \leq N$ , so long as  $X_i \neq X_j$  often enough to compute the median (Theil 1950; Sen 1968). This statistic estimates the slope  $\beta$  in the regression model  $Y \sim \beta X + \varepsilon$  (where  $E(\varepsilon) \neq 0$ ), in a more robust way than does the least squares method. If the regression errors  $\varepsilon$  are symmetric (or the  $X_i$  are ordered and symmetrically located, or  $N < 3$ ), then the estimator is unbiased (Wang and Yu 2005).

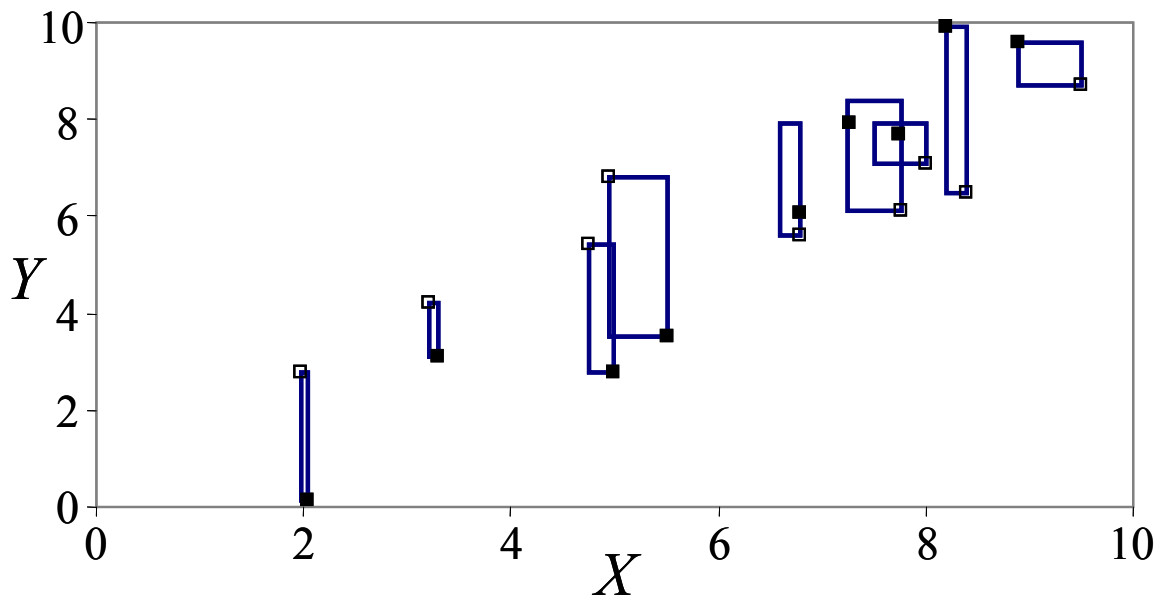
The Theil-Sen estimator can be generalized for interval data, but, like many other statistics, it cannot be sharply computed using the methods of naive interval analysis because of the repetition of uncertain quantities implicit in the expression for the statistic.

*Numerical example.* Suppose we have the following bivariate interval data:

$X$	$Y$
[4.95, 5.5]	[3.5, 6.8]
[8.9, 9.5]	[8.7, 9.6]
[7.25, 7.76]	[6.1, 8.4]
[4.75, 4.98]	[2.8, 5.4]
[6.6, 6.8]	[5.6, 7.9]
[8.2, 8.4]	[6.5, 9.9]
[1.975, 2.031]	[0.15, 2.80]

[3.2, 3.3]	[3.1, 4.21]
[7.5, 8.0]	[7.1, 7.9]

These data are plotted as the nine rectangles in Figure 30. Using Excel's Solver utility (Fylstra et al. 1998), the range of Theil-Sen slopes for these interval data was found to be [0.591, 1.628]. This interval represents the range of Theil-Sen slopes that can be obtained by any configuration of nine scalar data points constrained to lie inside their respective rectangles. The extremal configurations are also shown in the figure as open and closed squares. Note that computing the Theil-Sen slope using naive interval analysis—as the interval median of the 36 interval quotients of data pairs—yields [0.422, 4.727], which is wider than the best possible bounds on the Theil-Sen slopes. Although this interval is non-optimal, i.e., the lower bound is shallower and the upper bound is steeper than possible, its range in this case is still clearly positive, and in some cases the result of the naive interval calculation may be tight enough for practical concerns.



**Figure 30. Bivariate interval data (rectangles) and configurations of points that produce the smallest slope (open squares) and largest slope (filled squares) in Theil-Sen nonparametric linear regression.**

The interval statistics approaches to regression described in this section should be contrasted with statistical methods commonly employed to handle censoring in regression analyses. When only one of the two variables is left-censored, standard survival analysis methods are typically used in regression studies (Schmitt 1985; Feigelson and Nelson 1985; Isobe et al. 1986). Schmitt (1985) described a weighted least-squares regression method for data sets in which both the independent and dependent variables could include non-detects. Akritas et al. (1995) described the extension of the Theil-Sen nonparametric regression method for data with non-detects in both variables.

## 6 Computability of interval statistics

The table below summarizes the computability results for statistics of data sets containing intervals that have been established in this report and elsewhere (Ferson et al. 2002a,b; 2005a; Kreinovich and Longpré 2003; 2004; Kreinovich et al. 2003; 2004a,b; 2005a,b,c; Wu et al. 2003; Xiang 2006; Xiang et al. 2006; Dantsin et al. 2006; Xiang et al. 2007a). The column headings refer to exemplar problems. For instance, the computabilities in the Means column apply to all the various kinds of means described in section 4.2. The Variance column also applies to the calculation of standard deviation. The Breadths column applies to calculating upper and lower confidence limits, outlier detection, and even-ordered central moments. It also applies to percentiles and medians. The Covariance column also applies to calculating correlations.

<i>Kind of data set</i>	<i>Means</i>	<i>Variance</i>	<i>Breadths</i>	<i>Covariance</i>	<i>Odd central moments</i>
<i>No intersections</i>	$\Theta(N)$	$\Theta(N)$	$\Theta(N \log N)$	$\Theta(N^2)$	$\Theta(N^2)$
<i>Same precision</i>	$\Theta(N)$	$\Theta(N)$	$\Theta(N \log N)$	$\Theta(N^3)$	?
<i>Binned data</i>	$\Theta(N)$	$\Theta(N)$	$\Theta(N \log N)$	$\Theta(N^2)$	?
<i>Detect/non-detect</i>	$\Theta(N)$	$\Theta(N)$	$\Theta(N \log N)$	?	?
<i>Few intersections</i>	$\Theta(N)$	$\Theta(N \log N)$	$\Theta(N \log N)$	$\Theta(N^2)$	$\Theta(N^2)$
<i>No nesting</i>	$\Theta(N)$	$\Theta(N)$	$\Theta(N \log N)$	?	?
<i>Thin</i>	$\Theta(N)$	$\Theta(N)$	$\Theta(N)$	?	?
<i>Arrangeable</i>	$\Theta(N)$	$\Theta(N^D)$	$\Theta(N^D)$	?	?
<i>General</i>	$\Theta(N)$	NP-hard	NP-hard	NP-hard	?

The symbol  $N$  in the table denotes the sample size of the data set, and the tabled expressions are all “order of” estimates  $\Theta(\cdot)$ . These estimates exclude any fixed overhead associated with a calculation and focus on the per-sample processing cost. They do not reveal the absolute computational time required, but only the relative time as a function of sample size. The idea is to indicate how much more difficult the calculation becomes when there are a lot of samples to process. For comparison, the task of printing out  $N$  numbers has a computational difficulty on the order of  $N$ , that is,  $\Theta(N)$ . The task of sorting  $N$  numbers has a computational difficulty on the order of  $N \log(N)$ . Operations such as these are certainly feasible for reasonably small  $N$ , although they become more and more demanding as sample size increases. Other algorithms require computational effort that grows with the square or cube of the sample size. These are a lot more intensive when the sample size is large, but may still be practical in many cases. We see in the last row of the table that the computability is worst for the general case where we can make no assumptions about the widths of intervals or their patterns of overlap. In this situation, all the important statistics other than the mean are reported as NP-hard. NP-hard problems get harder much faster as  $N$  grows and are generally not feasible for large sample sizes. In practice, exact solutions may be computable only for data sets with a handful of values. However, the happy news of this report is that, for a wide variety of common cases summarized in the other lines of the table, the computabilities are much

less onerous. A question mark in the table denotes an open research question about computability.

The rows of the table refer to the kinds of interval data sets defined in section 3.5. Recall that in a no-intersections data set no two intervals may overlap each other (unless one is a degenerate, point value). In same-precision data, the intervals all have the same width. In binned data, the intervals are either identical (overlap perfectly), or don't overlap at all (or do so only at one point at most). In detect/non-detect data sets, the only nondegenerate intervals have the form  $[0, DL_i]$ . A data set with few intersections has fewer than  $K+1$  intersections among the intervals. Surprisingly, the number  $K$  doesn't play in the computability results. In a no-nesting data set, no interval is entirely within another. For arrangeable data, there are  $D$  subgroups each of which has no nesting.

The computabilities given in the table are theoretical, and their implications for practical computation should be well understood. The computabilities are order-of-magnitude calculations and only express *relative* costs of the algorithms. They reflect worst case times, neglecting the fact that the necessary calculation for a particular problem may be much less than suggested by the order given in the table. They also ignore all constant multipliers which determine the per-step computational cost as well as any overhead costs, which are usually important in practical considerations. For instance, as we explained at the end of section 4.5.1.2.4, an  $\Theta(N)$  algorithm is not necessarily better than an  $\Theta(N \log N)$  algorithm for a particular problem if the time required for each step of the former is a lot longer than for the latter. When the sample size is small or moderate, it may often be the case that the  $\Theta(N \log N)$  algorithm is faster than the  $\Theta(N)$  algorithm. Such considerations should be implemented in practical software intended to make these calculations automatically. Nevertheless, these computability estimates are important both theoretically and practically when the sample sizes become large. They inform us about how large sample sizes can practically be. It is likely that automated data collection will become more common in the future, so sample sizes may tend to become quite large if this is the case.

There may be considerable economies that arise because only some of the data are intervals. When  $H$  of the values are nondegenerate intervals and  $(N - H)$  are scalar point values, the computabilities improve according to this table.

<i>Kind of data set</i>	<i>Means</i>	<i>Variance</i>	<i>Breadths</i>	<i>Covariance</i>	<i>Odd central moments</i>
<i>No intersections</i>	$\Theta(N)$	$\Theta(N)$	$\Theta(N+H \log H)$	$\Theta(N + H^2)$	$\Theta(N + H^2)$
<i>Same precision</i>	$\Theta(N)$	$\Theta(N)$	$\Theta(N+H \log H)$	$\Theta(N H^2)$	?
<i>Binning data</i>	$\Theta(N)$	$\Theta(N)$	$\Theta(N+H \log H)$	$\Theta(N + H^2)$	?
<i>Detect/non-detect</i>	$\Theta(N)$	$\Theta(N)$	$\Theta(N+H \log H)$	?	?
<i>Few intersections</i>	$\Theta(N)$	$\Theta(N+H \log H)$	$\Theta(N+H \log H)$	$\Theta(N + H^2)$	$\Theta(N + H^2)$
<i>No nesting</i>	$\Theta(N)$	$\Theta(N)$	$\Theta(N+H \log H)$	?	?
<i>Thin</i>	$\Theta(N)$	$\Theta(N)$	$\Theta(N+H \log H)$	?	?
<i>Arrangeable</i>	$\Theta(N)$	$\Theta(N+H^D)$	$\Theta(N+H^D)$	?	?
<i>General</i>	$\Theta(N)$	NP-hard	NP-hard	NP-hard	?

In both of the above tables, the computabilities refer to algorithms designed to obtain *exact* results. For very large sample sizes, it may be necessary to abandon these methods and seek non-optimal approaches that can at least conservatively bound the statistics being sought. For instance, the jumping algorithms to compute bounds on the variance can produce results in only  $\Theta(N \log N)$  time, even in the general case where we can make no presumptions about the widths or overlapping among intervals. Such outer approximations that retain the guarantee associated with interval analyses could likewise be useful in estimating other statistics as well. Likewise, widening a few intervals in general interval data set could sometimes change it into a no-nesting, same-precision, binned or thin data set for which a convenient and fast algorithm is available. Although not epistemologically desirable, this might make calculations computationally practical and so might be part of a strategy for outer approximation.

The algorithms outlined in this report for important special cases, together with conservative bounding when those cannot be used, can make interval statistics practical for almost all applications. Only a few years ago, it might have been thought that calculating basic descriptive statistics for interval data sets would be computationally prohibitive. The surprising result summarized in this report is that, for many practically important cases, the calculations can be not only feasible but efficient.

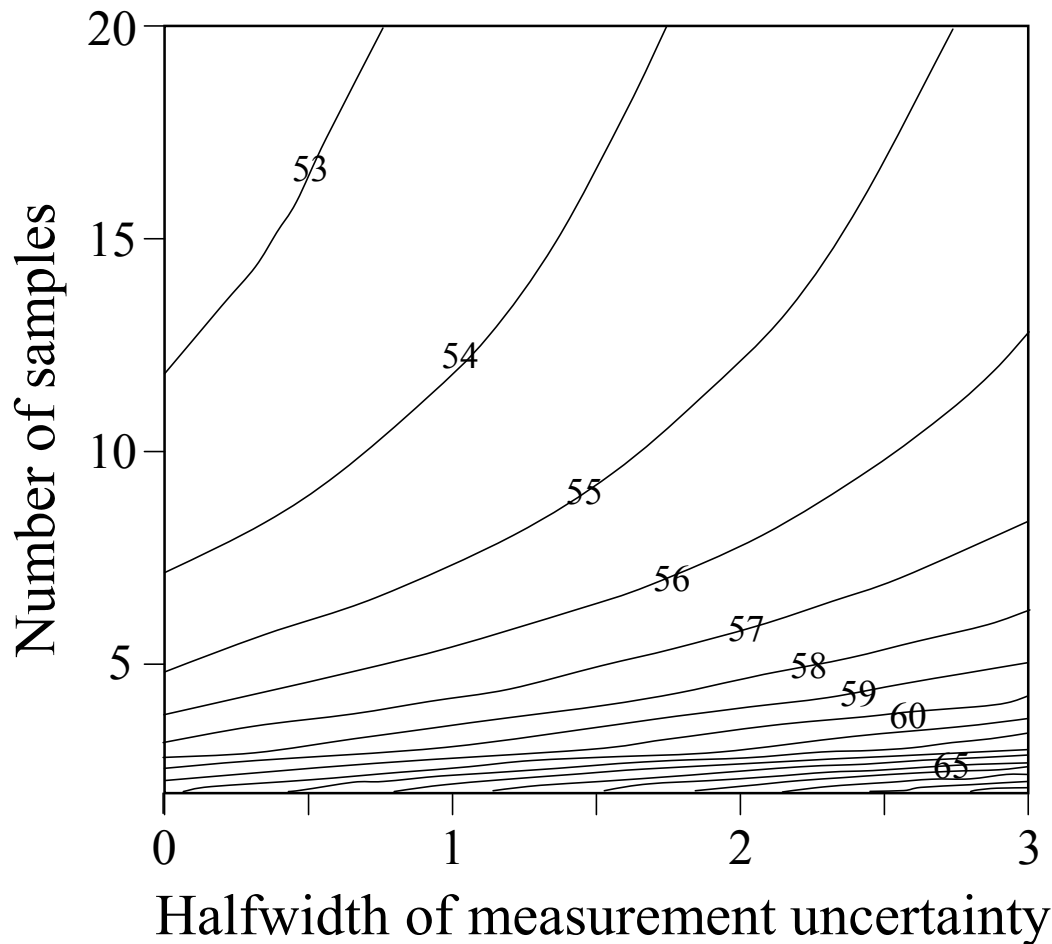
## 7 Tradeoff between sample size and precision

When describing a restaurant, Woody Allen once joked “It’s terrible. The food is bad and the portions are small”. The jest hinges of course on the notion that you would care about portion size when you wouldn’t want to eat the food in the first place. Data sets aren’t like meals. We certainly want the data to be good, but even if they are not good we’d still want them to be abundant. Scientists and engineers are starving for data, and we always want more. We’d like the data to be perfectly precise too, but we want them even if they come with unpalatably large imprecision. Likewise, we always want better data with greater precision, perhaps especially if few measurements can be taken.

Intuitively, given a fixed budget for empirical effort, one expects there to be a tradeoff between the precision of measurements and the number of them one can afford to collect. For instance, an empiricist might be able to spend a unit of additional resources to be devoted to measurement either to increase the number of samples to be collected, or to improve the precision of the individual samples already planned. Some practitioners apparently believe, however, that this tradeoff always favors increasing the number of samples over improving their precision. This belief is understandable given that most of the statistical literature of the last century has focused on assessing and projecting sampling uncertainty, and has in comparison neglected the problem of assessing and projecting measurement uncertainty. The belief is nevertheless mistaken, as can easily be shown by straightforward numerical examples. This section introduces one such example.

Consider the problem of estimating confidence limits on the mean from sparse and imprecise data. Such confidence limits are sometimes used as a conservative estimator of the mean that accounts for the sampling uncertainty associated with having made only a few measurements. The limits are affected by the sample size, but, if the calculation also accounts for the imprecision of the data values represented by intervals, they would also be affected by the measurement precision. Using the algorithms described in section 4.8 to compute confidence limits on the mean for interval data sets, Hajagos (2005) described a nonlinear tradeoff between precision and sample size as they affect the upper bound on the upper confidence limit. He conducted Monte Carlo simulation studies in which he varied both the precision and the sample size for data sets. His result is summarized in Figure 31, in which the abscissa gives the measurement precision for all data values and the ordinate gives the sample size of measurements all at that precision that can be collected given some fixed total cost for the empirical effort (assuming that both improving precision and increasing sample size have some cost). The isoclines give the values of the average resulting upper bound on the 95% upper confidence limit for the data set computed from random samples of the given sample size taken from the underlying population, each with the prescribed interval precision implied by the measurement uncertainty given on the abscissa. The numerical details are unimportant and depend on the choices Hajagos made about the underlying distribution in his simulations. What is interesting are the facts that (1) there is a tradeoff, and (2) it is nonlinear.



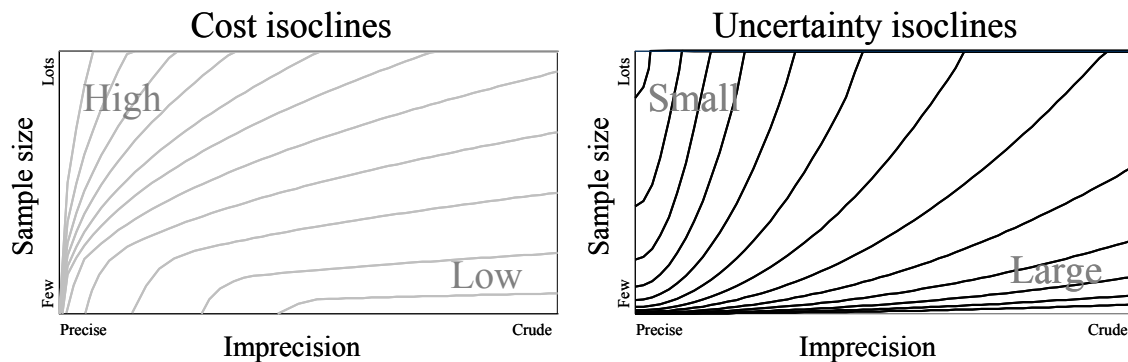


**Figure 31. Isoclines describing the tradeoff between measurement precision and sampling effort for the upper bound of an upper confidence limit. Redrawn from Hajagos (2005).**

In planning an empirical study from which an upper confidence limit will be estimated—and presuming that its value will have some engineering significance such as how robust a system design will have to be—this tradeoff between measurement precision and sample size will become centrally important. Studying it for a particular problem will allow engineers to determine whether they would be better off getting more or better data. The nonlinearity exemplified in the figure means the optimal investment of empirical resources between increasing sampling and improving precision depends on the quantitative details of the problem. An analyst who is interested in the upper bounds on the upper confidence limit (UCL, section 4.8), might plan an optimal empirical design in reference to figures such as this. The result can be understood as decisions about the lesser of two evils: sampling uncertainty versus measurement uncertainty. Sampling uncertainty arises from not having measured all members of a population and it is represented by confidence limits. Measurement uncertainty, specifically the uncertainty associated with making a finite-precision measurement, is represented by (non-statistical) interval ranges.

Let us briefly consider the implications of this tradeoff in the design of experiments. Suppose there are several measurement options available. We could pay \$10,000 for an exquisite measurement using the best, specially constructed apparatus and the most careful protocol executed by the world's foremost experimentalists. We could instead pay \$1,000 for a measurement that is merely excellent. Or we might choose a good-but-not-great measurement that costs only \$100. At the bottom of this barrel, we might also pay 10 bucks for what is admittedly a poor measurement. If we only have \$40 to spend, then our choice is to buy four of the poor measurements. If we are prepared to spend \$10K on the study, we could afford to buy the exquisite measurement, but it may not be true that we would want to spend it all in one place, as it were. We could instead buy 10 excellent, or a 100 good measurements. We might even prefer to buy 1000 poor measurements if the effect of sampling uncertainty is more important, i.e., has a worse effect on our ability to draw conclusions from the study than the measurement uncertainty from those poor measurements.

To study this problem quantitatively, we need to sketch the isoclines for cost and compare them to the uncertainty surface such as the one Hajagos computed (Figure 31). Suppose that Figure 32 shows cost isoclines in the left graph and the overall uncertainty isoclines for a statistic qualitatively similar to the upper UCL on the right graph. Costs are largest in the upper, left corner of the plane where there are lots of samples of high precision. Uncertainty is largest in the lower, right corner of the plane when there are measurements are crude and few. We want to balance costs with uncertainty so that neither is disproportionately large. The shapes of these isoclines are hypothetical, but they are similar to what might be expected in many real-world situations.



**Figure 32. Hypothetical isoclines for cost (left graph) and resulting uncertainty (right graph) for different experimental designs.**

Suppose that our experimental budget is small such that it corresponds to the lower gray cost isocline in Figure 33. The optimal precision for this budget is denoted by the lower filled circle. It says that we ought to have a small sample size of data with intermediate precision. If the project suddenly becomes more visible and the experimental budget is substantially increased, most project managers might plan to substantially increase sample size. Suppose that we can now afford a lot more of those samples as indicated by the empty circle in the figure directly above the first filled circle. This analysis suggests that collecting a lot more samples would be a decidedly bad way to spend the extra

money. Instead, we should follow the cost isocline selected by the empty circle to its uncertainty optimum, which would be the closed circle on the left. This denotes the optimum experimental design given that we have the substantially increased budget. This optimum suggests, however, that the extra money should mostly go to improving the precision of the measurements and only a little to collecting additional samples.

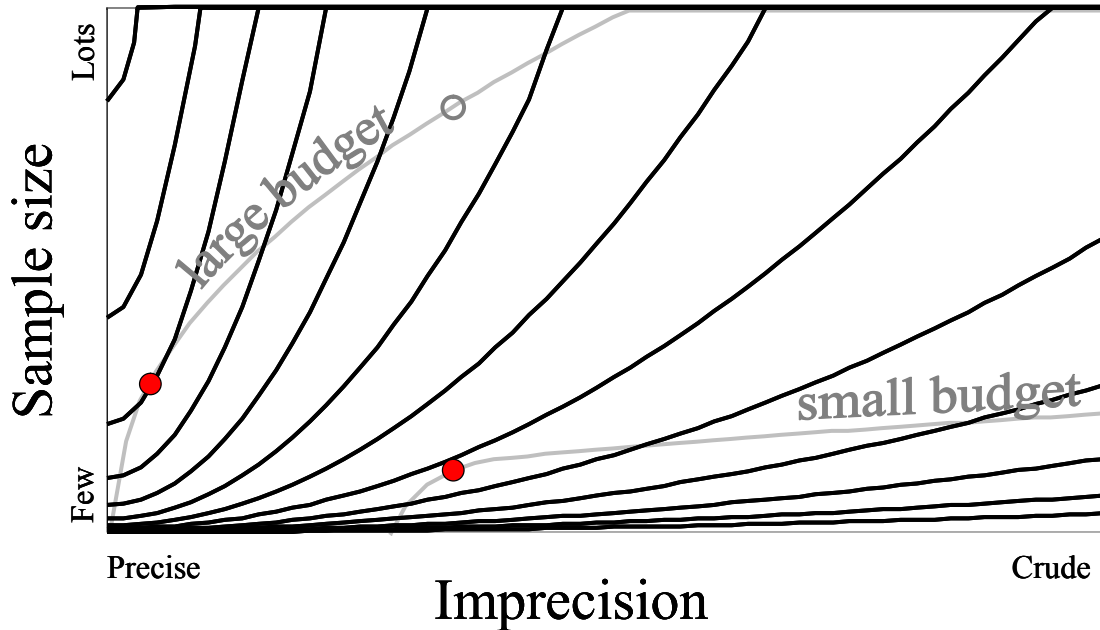


Figure 33. Optimum designs (filled circles) for each of two cost isoclines (gray) superimposed on the uncertainty isoclines (black).

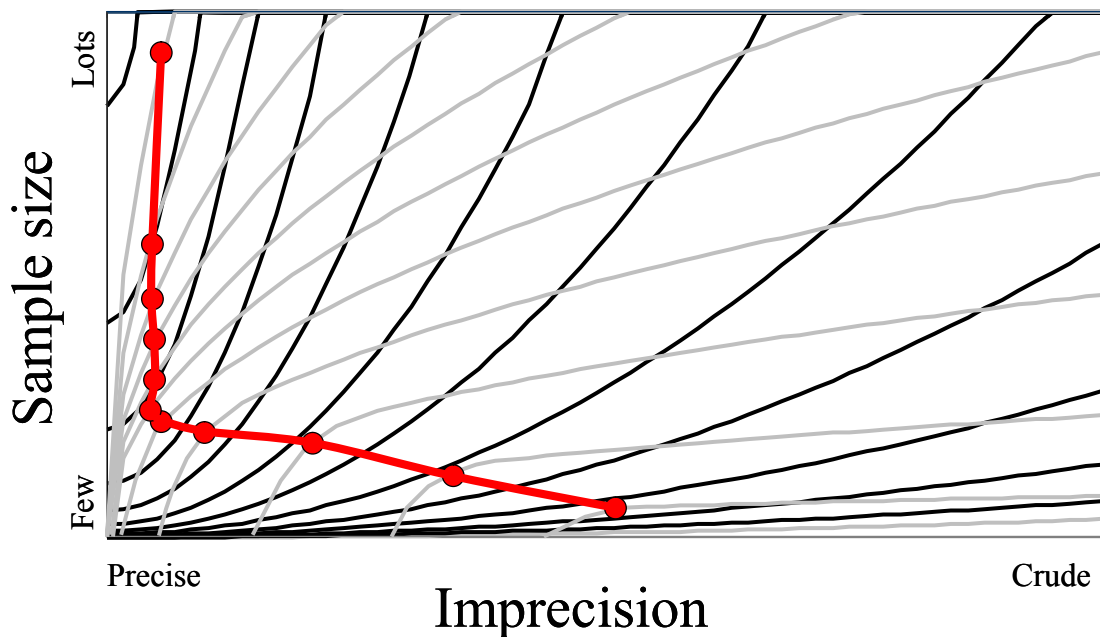


Figure 34. Line of best experimental designs from cheap (lower, right) to expensive (upper, left).

Figure 34 shows an L-shaped line of optimal experimental designs for many budget sizes, ranging from the very large at the topmost point on the L to the very small at the rightmost point of the bottom of the L. These points were obtained by optimizing uncertainty on each cost isocline, or equivalently, by optimizing cost on each uncertainty isocline. The L-shape of this line of optimal experimental designs is interesting. It suggests that when the budget is small (at the bottom of the L), an incremental increase in the budget should be devoted mostly to increasing precision, whereas when the budget is above the corner of the L, the increment should be devoted mostly to increasing sample size. Thus, it appears that the designs are either in what might be called a “more-sampling” domain on the left side of the L, or in a “more-precision” domain on the bottom of the L. Of course, this particular shape depends on the local numerical details of the problem, which were hypothetical. Nevertheless, it is perhaps reasonable to anticipate that other, similar problems might also display both more-sampling and more-precision domains.

The isoclines will depend on the details of the empirical methods employed. Naturally, they don’t have to look just like these examples in any particular situation, although they often will be similar because of common economies of scales. For instance, measurement laboratories frequently offer their customers volume discounts. Statistical confidence calculations likewise have stereotypical asymptotic behaviors of diminishing returns, so the first few samples tend to tell us more than the last few samples do. Analysts should be able to construct their own isoclines for a particular project given specific cost schedules and characteristics of the experimental design.

The tradeoff that has been discussed here is that between precision and sample size, which is concerned with how one’s budget for empirical work will be spent. Analysts are often also interested in another tradeoff as to how big the empirical budget should be. This tradeoff problem could be studied in a wider context that allows the comparison to be expressed purely in monetary units. For example, by translating the overall uncertainty in the statistical summary into dollars—small uncertainty is cheaper because, say, it would allow for a more efficient engineering design—one can add the up-front empirical costs of a project to the resulting long-term costs of potential overdesign and then minimize the total project cost. This would involve “value of information” considerations and could permit a reasoned inference about how much investment in empirical effort is appropriate for a project. Such considerations ought to include, and distinguish, epistemic and aleatory uncertainties.

Of course such tradeoffs between precision and sample size have been appreciated by others, who have occasionally found that devoting resources to improving precision may be preferable to collecting more samples in some situations. Using the confidence limits described by Horowitz and Manski (2000), Manski (2003, page 20f) described the computation of confidence intervals of a statistic for data with missing values and found that the incertitude arising from the missing data was much larger than the uncertainty resulting from incomplete sampling at a specified level of confidence. This result occurred unless the sample size was very small or there were very few missing values. Findings like these contradict the prevailing intuition among statisticians that it’s better to

get more data than to improve measurement precision. Engineers have not traditionally subscribed to this idea, but they are being exposed to it more and more as statistical training becomes more pervasive in engineering.

Many large or publicly funded empirical studies today are governed by quality assurance guidelines (e.g., EPA 2000). Under these guidelines, measurement quality is a yes-no consideration. A measurement or a data set either meets quality criteria and is acceptable, or it does not meet these criteria and is deemed unacceptable and is excluded from the study. Although these quality criteria are context-dependent and will usually differ for different studies, the assessment of data quality is always a binary decision. This yes-no thinking extends from individual measurements to entire projects. For example, guidelines might prescribe that data from a particular laboratory or study can be used if at least 90% of the measurements have no larger than  $\pm 75\%$  uncertainty. Yet, empiricists have a strong intuition that measurement quality can vary continuously (as is evidenced by the fact that the phrase “ $\pm 75\%$  uncertainty” itself is meaningful). And one might expect that discarding even poor data would be an anathema to an empiricist if it contains any extractable information. It seems to us that a clearly better and more flexible strategy would be to keep the data while explicitly acknowledging their imprecision with some appropriate discounting. This does not seem to be practiced in data quality guidelines as commonly defined, even though it would be possible within a traditional statistical approach using a weighting scheme. Interval statistics allows still more expressiveness in accounting for incertitude in data. It seems to us that data quality criteria as currently prescribed could be improved in obvious but profound ways by judicious use of interval statistics.

## 8 Methods for handling measurement uncertainty

The interval statistics approach to modeling epistemic uncertainty associated with empirical data that is reviewed in this report can be applied to the important problem of evaluating, representing and propagating measurement uncertainty. It constitutes a possible alternative to or generalization of current approaches (Coleman and Steele 1999; Bevington and Robinson 1992; Taylor and Kuyatt 1994; Carroll et al. 1995; Dieck 1997) which require sometimes untenable assumptions and cannot really capture interval uncertainty. This approach will be of limited interest for the lucky metrologists who can properly neglect resolution uncertainty as a very small part of their overall uncertainty and don't encounter censored or missing data. For others, however, it may be much more useful, especially at the frontiers of metrology, where new measuring technologies are employed and the limits of measurability are being stretched. Section 8.1 will briefly review the standard approach to representing and propagating measurement error used by many national and international standards and metrology institutes and recommended by many scientific and engineering societies. Section 8.2 applies the methods of interval statistics to the standard approach for representing and propagating measurement uncertainty. Section 8.3 discusses how interval statistics and the standard approach might be hybridized to have the advantages of both approaches in allowing for substantial interval uncertainty and yet also acknowledging that some errors may be larger than the nominal bounds given by the empiricist. Section 8.4 will consider the question of whether there is any demonstrable need to do things differently in handling measurement uncertainty.

### 8.1 *The standard approach*

A measurement result is complete only if it is accompanied by a quantitative expression of its uncertainty. The uncertainty is needed to judge whether the result is adequate for its intended purpose and whether it is consistent with other similar results. Over the last few decades a broad consensus has emerged about how measurement error should be quantified and incorporated into subsequent calculations. In 1977, the International Committee for Weights and Measures (CIPM), which is the highest authority in the field of metrology, asked the International Bureau of Weights and Measures (BIPM) to study the prospects of reaching an international agreement on expressing uncertainty in measurements in collaboration with various national metrology institutes. The substantial result was the report of Kaarls (1981), the main recommendations of which were approved in 1981 by the CIPM. The appendix on page 133 of this report reproduces these recommendations. In 1985, CIPM asked the International Organization for Standardization (ISO) to develop guidance based on the BIPM recommendations, resulting in an elaborated document (ISO 1993) representing the international view of how to express uncertainty in measurements. This guidance was succinctly presented in a technical note from the National Institute of Standards and Technology (NIST) (Taylor and Kuyatt 1994). Dieck (1997) also provided a tutorial introduction to the subject. Since 1997, the Joint Committee on Guides in Metrology (JCGM) has assumed

responsibility for the guidance and has recently released a supplement on how Monte Carlo simulation can be used to extend the applicability of the approach (JCGM 2006).

The language of the guidance abandons the familiar ideas of random and systematic errors, and talks instead about Type A and Type B evaluations, the difference between which is related to how the uncertainties are estimated. The evaluation of uncertainty by a statistical analysis of a series of observations is called a Type A evaluation of uncertainty. An evaluation of uncertainty by any other means is called a Type B evaluation. The phrase “other means” typically refers to what might be called scientific judgment. Both Type A and Type B evaluations of uncertainty are expressed in terms of probability distributions summarized by a standard deviation of the mean. Typically, the distribution is presumed to be a normal (Gaussian) distribution, but sometimes a uniform (rectangular) or triangular (Simpson) distribution, or other distribution shape may be assumed. However, even if the uncertainty is initially presumed to have a uniform distribution, once it has been summarized with a standard deviation of the mean, it is subsequently treated as though it were normally distributed when used in calculations. These probability distributions represent the uncertainty about a fixed quantity that would be associated with repeated measurements of it. For instance, a statistical analysis of  $N$  repeated direct measurements  $X_i$  of some quantity would be summarized as the best estimate

**Expression 1** 
$$Y = \frac{1}{N} \sum_{i=1}^N X_i$$

called the “measurement result”, together with its uncertainty

**Expression 2** 
$$U(Y) = \sqrt{\frac{1}{N(N-1)} \sum_{i=1}^N \left( X_i - \frac{1}{N} \sum_{j=1}^N X_j \right)^2},$$

which is called the “standard uncertainty” associated with the measurement result  $Y$ . This uncertainty is known in statistics as the standard error\* associated with the collection of measurements. Note that this expression becomes smaller and smaller as the sample size  $N$  gets larger and larger. This is a central feature of this approach to modeling measurement uncertainty. Because of its obvious connections with the central limit theorem for normal distributions, we might call it the ‘centrality feature’ of the standard approach. It represents the idea that, with sufficiently many estimates of a quantity, the uncertainty about that quantity can be made as small as desired. In contrast, an approach based on interval statistics can allow for situations in which incertitude represents a fundamental constraint on the possible precision of an estimate. Once we have reached this constraint, no matter how many more estimates are taken, the aggregate estimate cannot get any better. The standard approach recognizes no such constraint.

---

\*This quantity is also commonly called a standard deviation, because it is the standard deviation of a distribution of mean values (all of the same sample size  $N$ ), but this usage can be confusing because it may be ambiguous as to whether it refers to the dispersion of the original  $X$  values or of the means of  $X$  values.

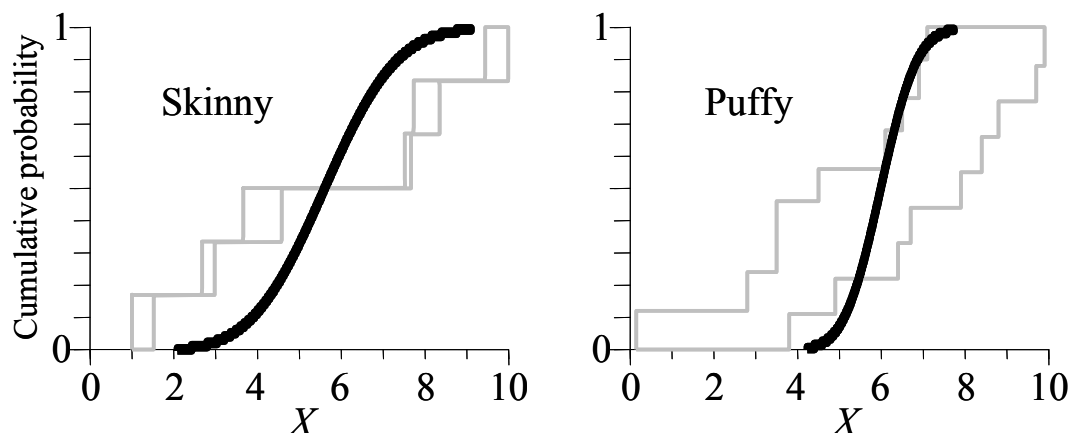
NIST uses the plus-or-minus symbol  $\pm$  between the measurement result and its standard uncertainty to denote the normal distribution model of the measurement and its uncertainty. For instance, the expression  $0.254 \pm 0.011$  denotes a normal distribution  $N(0.254, 0.011)$  with mean 0.254 and standard deviation 0.011 which is the model for the uncertainty of the measurement result. This notation is somewhat at odds with other common uses of the  $\pm$  symbol to denote a pair of values of differing sign or a range of values around a central point. One must be careful to keep in mind that the plus-minus notation between the measurement result and its standard uncertainty denotes a *probability distribution*, rather than an interval\* or a pair of scalar values. NIST also suggests a “concise notation” in which the standard uncertainty is expressed parenthetically as mantissa digits whose power is specified positionally by the last digits of the best estimate. Thus, the concise notation 0.254(11) is equivalent to the normal distribution  $0.254 \pm 0.011$ .

*Numerical examples.* As an example of a Type A evaluation, we apply Expressions 1 and 2 to the skinny and puffy data sets. The skinny and puffy data were given on page 28 and plotted in Figure 6. The most likely treatment by analysts would simply ignore the express uncertainty of the intervals (although we consider another way these data sets might be handled within the ISO/NIST approach below). Replacing the intervals in the skinny data set by their midpoints yields the values 1.26, 2.83, 7.595, 8.04, 9.715, and 4.12. The mean of these six scalars is 5.593 and their standard error is 1.361, with ordinary rounding. Thus the model of measurement uncertainty is the normal distribution with mean 5.593 and standard error 1.361, which can be written as  $5.593 \pm 1.361$  or as 5.6(14). The scalar midpoints of the puffy data are 4.95, 7.85, 7.25, 4.75, 6.6, 8.2, 1.975, 4.7, 7.5. These nine values lead to a mean of 5.975 and a standard error of 0.6768, and, with ordinary rounding, the normal distribution written as either  $5.98 \pm 0.68$  or 5.98(68). These distributions are plotted in Figure 35 as black curves against the empirical distributions originally plotted in Figure 6 shown in gray. The black distributions are steep compared to the gray distributions for two reasons. First, the uncertainties in the interval data are ignored because only the midpoints of the intervals are used, and, second, the standard error (standard deviation of the mean) is the dispersion of interest, not the standard deviation of the data themselves.

---

\*Confusingly, NIST also uses the  $\pm$  symbol between the measurement result and an “expanded uncertainty”, which is the standard uncertainty multiplied by a coverage factor such as 2, 3 or some critical value from the Student  $t$  distribution that produces something very similar to a confidence interval around the measurement result. When the coverage factor is 2 or 3, the level of confidence is taken to be about 95.4% or 99.7%, as would be expected assuming normality. When the coverage factor is one and the number following the  $\pm$  sign is the standard uncertainty, the coverage probability is presumed to be about 68.2%. The usage, it could be argued, denotes an interval, although the distributional qualities of the model invoked are always paramount. NIST’s guidance on reporting uncertainty requires the term following the  $\pm$  symbol to be explicitly named and the value of the coverage factor, if used, to be given.





**Figure 35. Standard ISO/NIST models (black curves) of measurement uncertainty neglecting incertitude for the skinny and puffy data, superimposed on their empirical distributions (gray).**

In the case of a Type B evaluation, the standard uncertainty can be estimated from calibration data or other previous data, or the manufacturer's specifications for the measuring instrument, or the analyst's past experience with the device or similar devices. It can also be estimated by backcalculation from a previously reported confidence interval assuming normality, or even by backcalculating from a reported range (that is not described as a confidence interval) by assuming a coverage probability and a distribution shape and treating the range specified as though it were a confidence interval.

The quantity being measured, the measurand, is not always measured directly. Sometimes it is measured indirectly as a function of other measured quantities. For instance, we might use Ohm's law to indirectly measure an electrical current from directly measured values for resistance and voltage, or we might indirectly measure the mass of water in a pool from direct measurements of the length, width, and height of the pool and the temperature and salinity of the water. In these cases, and, indeed, in most metrological work, we must compute the measurement from other quantities. The standard approach recommended by ISO and NIST specifies that these calculations should propagate measurement uncertainties as well as the central values. If, for example, the equation that relates an indirectly measured value  $Z$  to other measured quantities of  $Y_1, Y_2, \dots, Y_L$  is

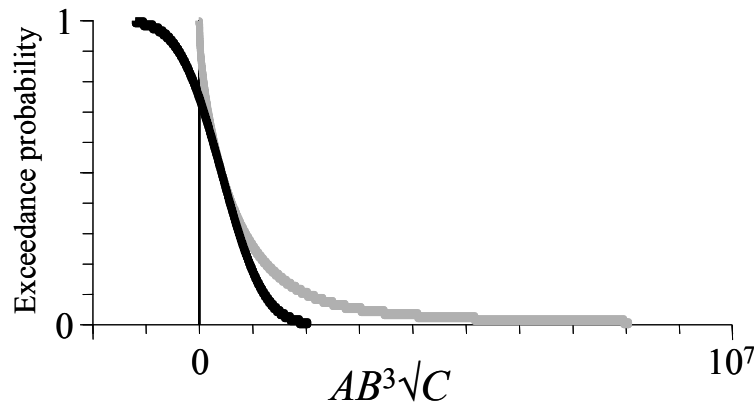
$$Z = F(Y_1, Y_2, \dots, Y_L),$$

then the function  $F$  should be applied to the best estimates of  $Y_1, Y_2, \dots, Y_L$ , which are perhaps obtained from Expression 1, to obtain the best estimate of  $Z$ . The uncertainty associated with this calculation is the weighted quadrature sum of the input uncertainties

**Expression 3** 
$$U(Z) = \sqrt{\left[\frac{\partial F}{\partial Y_1}\right]^2 U^2(Y_1) + \left[\frac{\partial F}{\partial Y_2}\right]^2 U^2(Y_2) + \dots + \left[\frac{\partial F}{\partial Y_L}\right]^2 U^2(Y_L)}$$

where  $U(Y_i)$  is the standard uncertainty associated with  $Y_i$ , perhaps estimated from Expression 2. When the function  $F$  is simple addition, Expression 3 reduces to the root sum of squares of the uncertainties. When  $F$  is averaging as in Expression 1, Expression 3 reduces to Expression 2 if we allow the uncertainty of a single value to be the standard deviation of the distribution from which it's drawn. This weighted quadrature sum formula is called the “law of propagation of uncertainty” for independent inputs.

The formula is exact when  $Y_1, Y_2, \dots, Y_L$  are independent and normally distributed and  $F$  is linear. It is also reasonable as the first terms in a Taylor approximation in other cases when the  $U(Y_i)$  are small in magnitude. The approximation suffers when uncertainties are large and the model nonlinear, however. For example, suppose  $F(A, B, C) = AB^3\sqrt{C}$ , where  $A$  is modeled by  $12 \pm 1.2$ ,  $B$  is  $34 \pm 17$ , and  $C$  is  $0.78 \pm 0.195$ . The central value is  $F(12, 34, 0.78) = 12 \times 34^3 \times 0.78^{1/2} \approx 416,548$ , and its standard uncertainty is  $\sqrt{((\partial F/\partial A)^2 U(A)^2 + (\partial F/\partial B)^2 U(B)^2 + (\partial F/\partial C)^2 U(C)^2)} = ((34^3 \times 0.78^{1/2} \times 1.2)^2 + (3 \times 12 \times 0.78^{1/2} \times 34^2 \times 17)^2 + (1/2 \times 12 \times 34^3 \times 0.78^{-1/2} \times 0.195)^2)^{1/2} \approx 628,370$ . The model of uncertainty in the measurement result is thus the normal distribution  $416,548 \pm 628,370$ , which is depicted in Figure 36 as the black curve in complementary cumulative form. Yet, if the variables actually have the distributions as modeled, that is,  $A \sim N(12, 1.2)$ ,  $B \sim N(34, 17)$ , and  $C \sim N(0.78, 0.195)$ , then the actual distribution of the random function  $AB^3\sqrt{C}$ , which can easily be computed with Monte Carlo or other methods, is that depicted in Figure 36 as the gray curve. The actual distribution is obviously not normal in shape. It has a standard deviation almost 40% larger than the ISO/NIST approximation suggests. Only about 2% of the distribution's values are negative, rather than a quarter of them as suggested by the approximation. Moreover, the right tail of the actual distribution is severely underestimated. The ISO guidance has recently been supplemented by JCGM (2006) which discusses how Monte Carlo methods can be used in such situations.



**Figure 36. Poor approximation by ISO/NIST model (black) of actual uncertainty distribution (gray) when function is nonlinear and uncertainties are large.**

Sometimes uncertainty sources are not independent of one another. For instance, if the same device, say an optical range finder, is used to measure the length and width of a rectangular area, then it is likely that the errors, and thus the uncertainties, are highly correlated with each other. For such cases, the standard approach for handling

measurement uncertainty generalizes Expression 3 to take account of the covariances, or, equivalently, the correlations among the inputs (ISO 1993; Taylor and Kuyatt 1994). The covariances are sometimes called “mutual uncertainties”. The standard approach allows a user to specify any\* correlation between different measurement sources in the uncertainty budget. JCGM (2006) elaborates the guidance on handling correlations within Monte Carlo simulations, although it is perhaps misleading when it enthuses that “the implementation of the [Monte Carlo method] for input quantities that are correlated is no more complicated than when the input quantities are uncorrelated” (ibid., section 9.4.3.1.3). The guidelines envision an analyst specifying zero correlations for independent sources and positive or negative values from the permissible range  $[-1, +1]$  for sources that may be somehow related to each other physically. Perfectly dependent sources would be assigned +1, and perfectly negatively dependent sources would be assigned  $-1$ . The analyst decides which correlations to use in any particular case.

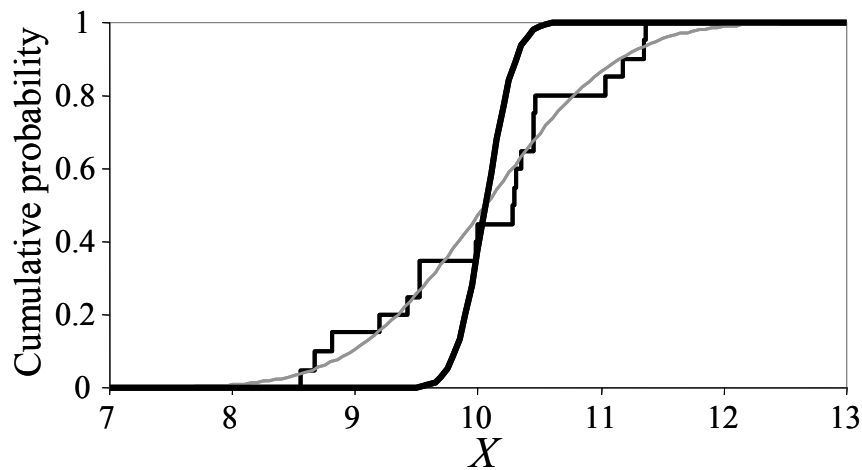
Although it is not expressly stated in the guidance, the variability represented by the standard uncertainty is the mean’s variability among several hypothetical studies involving the measurement, or perhaps among different scientists making the measurements. The measurand is presumed not to have any intrinsic variability itself (ISO 1993, sections 1.2 and 3.1.3). This is appropriate, or at least tenable, for many physical quantities such as the speed of light in a vacuum, Planck length, and the Newtonian gravitational constant. As a result of this focus, the ISO/NIST approach uses tight normal distributions (or sometimes rescaled  $t$ -distributions) to represent the imprecision of the measurement, whether it is a direct measurement or actually a calculation, and whether it is expressed as variability of repeated samplings or in some other, possibly subjective form. The distributions are ‘tight’ in the sense that they represent sampling variability among *means* of measurements, rather than the sampling variability of the measurements themselves. For example, suppose 20 measurements of a quantity have been made and the empirical distribution function of the results is shown in Figure 37 as the step function. The ISO/NIST model of the measurement imprecision is the tight normal distribution shown as the thick black curve. For comparison, the normal distribution fitted to the measurements themselves (by the method of moments) is shown as the thin gray curve. The difference is due to the ISO/NIST approach’s use of the standard error, rather than the standard deviation as the statistic of uncertainty. Because the standard error is obtained by dividing the standard deviation by the square root of the sample size, the ISO/NIST model of uncertainty is *always* tighter than the actual data. When the sample size increases, it becomes hard to distinguish the empirical distribution from the gray curve, while the black curve used by ISO/NIST gets steeper and steeper, i.e., more and more like a spike equivalent to a point value without any uncertainty at all.

---

\*It is apparently even possible to specify contradictory correlations between variables  $A$ ,  $B$  and  $C$  like this:

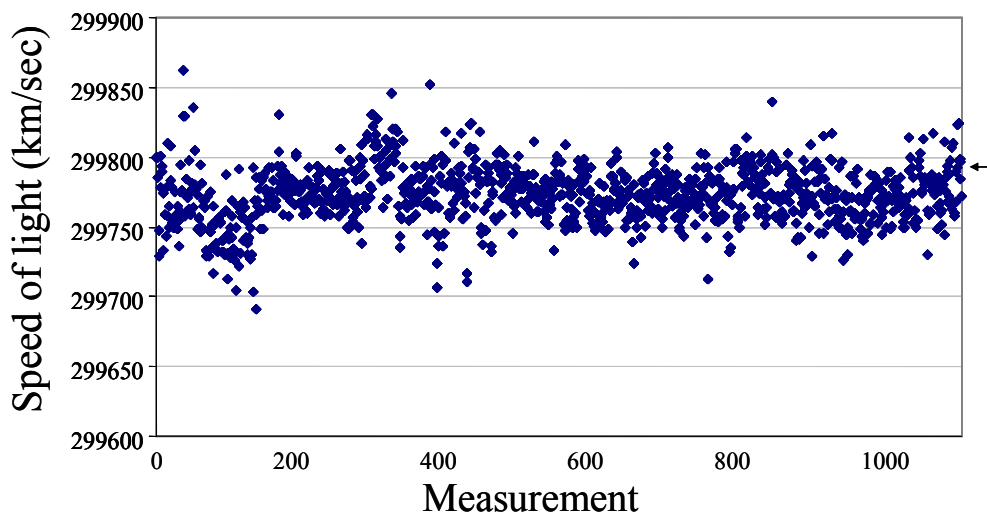
	$A$	$B$	$C$
$A$	+1	+1	$-1$
$B$	+1	+1	+1
$C$	$-1$	+1	+1

This matrix is symmetric and has permissible values (between  $-1$  and  $+1$ ), but it is not positive semi-definite and therefore cannot be a correlation matrix (see Ferson et al. 2004a). This oversight was corrected in JCGM (2006), which recommended such matrices be “repaired” by modifying them to be positive semi-definite.



**Figure 37. ISO/NIST model of measurement uncertainty (thick black curve) compared to the data's empirical distribution (step function) and fitted normal distribution (gray).**

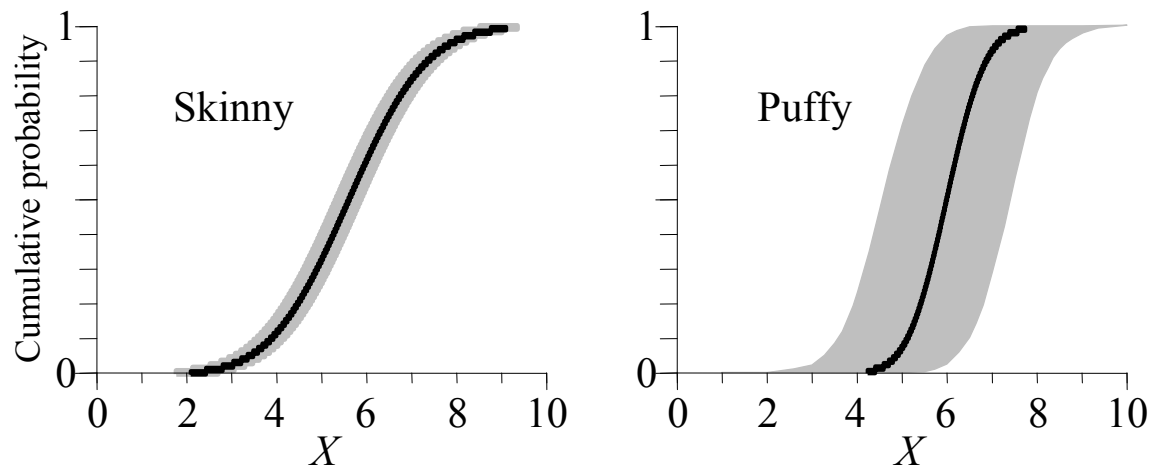
One might argue that this is a silly extrapolation and that no one collects so many repeated measurements of exactly the same quantity. But, whether or not it actually happens in fact, the asymptotic behavior is relevant in deciding whether a method for characterizing uncertainty is sensible or not. In fact, repeating measurements is exactly what empiricists are encouraged to do in order to minimize their uncertainty. Michelson and his colleagues (1935) famously conducted over a thousand separate determinations of the speed of light in 233 series over two years in Pasadena. Figure 38 displays their results, which apparently they only stopped collecting because the cows actually came home. It is worth noting, although it is hardly surprising, that the arithmetic mean of all of their estimates (299,774 km/s, with average deviation of 11 km/s) does not approach the currently accepted value of the speed of light, which is indicated by a small arrow on the right side of the figure.



**Figure 38. Repeated measurements of the speed of light by Michelson et al. (1935). Data values obtained from Cutler (2001b).**

## 8.2 Using interval statistics within the standard approach

The methods of interval statistics can be immediately employed within the ISO/NIST approach so that it can account for incertitude. Figure 39 depicts (in black) the traditional ISO/NIST models ignoring incertitude of the skinny and puffy data sets, which were already shown in Figure 35, together with analogous models (in gray) based on interval statistics accounting for incertitude. Following the ISO/NIST guidance (ISO 1993; Taylor and Kuyatt 1994), these models represent normal distributions fitted by the method of matching moments, but now using the techniques of interval statistics as described in section 4.9.1.



**Figure 39. Standard ISO/NIST models (black) for the uncertainty of estimates from the skinny and puffy data sets, together with their interval-statistics analogs that account for incertitude.**

The associated numerical summaries of the two data sets are given in the tables below. The estimates in the first column of the tables are the means of the data sets. The ISO/NIST calculations use Expression 1 with midpoints of the data intervals, and the interval statistics calculations use the generalization of Expression 1 described in section 4.2.1 for interval data. The standard uncertainties in the second column of the tables are the standard errors, computed with Expression 2 or its interval generalization described in section 4.5.2. The coverage intervals in the last column of the tables are computed from the estimate plus or minus  $k$  times its respective standard uncertainty. The interval statistics calculation uses the algorithm described in section 4.8.2. These coverage intervals were based on  $k = 2$ , which produces a probabilistically symmetric coverage interval with an associated coverage probability (or level of confidence) of approximately 95%, assuming the underlying distribution is normal. According to the ISO/NIST guidance (Taylor and Kuyatt 1994, section 5.4), the values of the measurands are believed to lie within these intervals with a level of confidence of approximately 95%. All interval statistics are reported with outward-directed rounding, but intervals from traditional ISO/NIST calculations are reported with ordinary rounding.

*Skinny*

	<i>Estimate</i>	<i>Uncertainty</i>	<i>Coverage interval</i>
<i>ISO/NIST</i>	5.593	1.361	[2.871, 8.315]
<i>Interval statistics</i>	[5.338, 5.849]	[1.258, 1.467]	[2.583, 8.591]

*Puffy*

	<i>Estimate</i>	<i>Uncertainty</i>	<i>Coverage interval</i>
<i>ISO/NIST</i>	5.975	0.677	[4.621, 7.329]
<i>Interval statistics</i>	[4.561, 7.389]	[0.338, 1.172]	[3.018, 9.004]

The coverage intervals computed above assume that the standard uncertainties have negligible uncertainty themselves. Because the sample sizes are rather low in these two data sets, that may be an untenable assumption. Some analysts in this circumstance might wish to use the Student  $t$ -correction to account for the small sample sizes. This correction assigns the coverage factor  $k$  to the critical value for a stipulated level of confidence of the two-sided  $t$ -statistic having degrees of freedom one less than the sample size. Thus, the correction would use  $k = 2.5706$  for the skinny data set and  $k = 2.3060$  for the puffy data set. The results after the Student  $t$ -correction represent statistical confidence intervals. The ISO/NIST calculation, which ignores incertitude, then leads to the 95% coverage interval of [2.095, 9.092] for the skinny data set and [4.414, 7.536] for the puffy data set. The parallel interval statistics calculation again uses the algorithm described in section 4.8.2 to compute the outer bounds on the two-sided normal 95% upper and lower confidence limits on mean. The envelope of these outer bounds, expressed with outward-directed rounding, is [2.725, 9.314] for the puffy data and [1.771, 9.401] for the skinny data. They are so wide mostly because there are so few sample values but, in the case of the puffy data, also because their incertitude is so large. These confidence intervals can be directly compared to the distribution-free confidence limits on the means for the skinny and puffy data in Figure 18. All three sets of intervals are summarized in the following table.

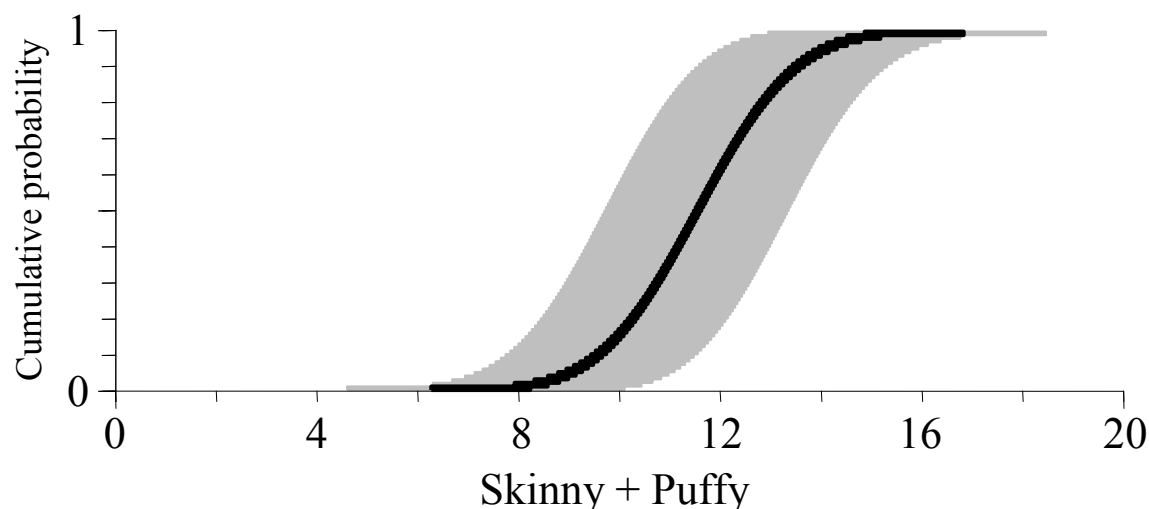
	<i>Skinny</i>	<i>Puffy</i>
<i>ISO/NIST</i>	[ 2.09, 9.09]	[ 4.41, 7.54]
<i>Normal with incertitude</i>	[ 1.77, 9.41]	[ 2.72, 9.32]
<i>Distribution-free</i>	[ 1.16, 14.70]	[ 1.70, 13.79]

For the interval statistics results in the second and third rows, we used outward-directed rounding, but for the ISO/NIST coverage intervals in the first row, ordinary rounding was used to conform with customary practice. The distribution-free intervals are widest, which is not surprising because they account for incertitude in the data and do not make use of the normality assumption. The ISO/NIST intervals are the tightest because they use normality and neglect the incertitude. The intervals in the middle row assume the underlying distributions are normal but account for the incertitude in the data sets.

Comparisons with the intervals in Figure 17 are more difficult because that figure characterizes the uncertainty of only the upper confidence limits, and also because it was based on a *one-sided* Student  $t$ -correction. It is interesting and perhaps surprising that the

configuration of scalar values within the respective data intervals that extremizes the endpoints of coverage and confidence limits may depend on whether a one- or two-sided limit is computed. This means that one cannot easily recompute coverage intervals or confidence intervals for different coverage probabilities or levels of confidence by simply altering the value of the  $k$ . In principle, each calculation requires a separate search for the extremizing configuration of scalars within the intervals. In practice, however, the actual numerical differences between the results may commonly be rather small in magnitude.

The objects in Figure 39 would typically be characterizations of the uncertainties of inputs to mathematical functions through which these uncertainties must be propagated. Suppose, for instance, that we are interested in the sum of the two quantities estimated from the skinny and puffy data sets. Figure 40 depicts the corresponding ISO/NIST estimate (in black) of this sum, together with its interval statistics analog (in gray). The black distribution was computed using Expression 3. Because the mathematical function is simple addition, the partials degenerate to unity. The same distribution could also have been computed using straightforward Monte Carlo simulation as described in JCGM (2006). The gray p-box in Figure 40 was obtained by straightforward calculations described and implemented by Williamson and Downs (1990; Berleant 1993; 1996; Ferson and Ginzburg 1996; Berleant and Cheng 1998; Ferson et al. 2003; Ferson and Hajagos 2004). The black distribution characterizes the estimate of the sum  $5.593 + 5.975 = 11.568$  with combined uncertainty  $\sqrt{1.361^2 + 0.6768^2} = 1.520$ . The coverage interval based on  $k = 2$  is  $[8.528, 14.608]$ , which should include the actual sum with a level of confidence of about 95%. In contrast, the analogous approach using interval statistics suggests a mean for the sum of  $[5.338, 5.849] + [4.56, 7.39] = [9.898, 13.239]$ , a combined uncertainty of  $\sqrt{[1.258, 1.467]^2 + [0.338, 1.172]^2} = [1.302, 1.878]$  with outward-directed rounding. A coverage interval with coverage probability 95% can be computed using elementary interval analysis as described in section 3.1, but reading directly from Figure 40 (i.e., the 2.5<sup>th</sup> percentile from the left bound and the 97.5<sup>th</sup> percentile from the right bound) gives the slightly tighter coverage interval of  $[6.5, 16.6]$ . Note that this is more than 65% wider than the ISO/NIST coverage interval that neglects incertitude. This example shows that the standard approach yields a starkly different picture of the uncertainty of the result compared to the result obtained from an approach based on interval statistics. In our view, the standard ISO/NIST answer suggests the overall uncertainty is smaller than it actually is on account of the incertitude. Nothing about the range or dispersion (slope) of the ISO/NIST distribution tells us anything about the breadth of the incertitude revealed in the interval statistics results. Thus, without an analysis that specifically propagates it, this information about the reliability of the estimate would be lost.



**Figure 40. Propagation of uncertainty about the skinny and puffy data sets to the sum under the ISO/NIST scheme (black) and via interval statistics (gray) assuming independence.**

The black distributions in Figure 39 and Figure 40 were based on calculations that completely ignore the incertitude in the data. Could a motivated analyst account for this incertitude using the ISO/NIST approach without interval statistics? The guidance (ISO 1993, section F.2.3.3; Taylor and Kuyatt 1994, section 4.6) prescribes that an interval such as we consider in this report should be treated as a uniform distribution so that, in particular if the interval datum is  $[\underline{x}, \bar{x}]$ , then the measurement result should be taken to be  $(\underline{x} + \bar{x}) / 2$ , with standard uncertainty  $U = (\bar{x} - \underline{x}) / \sqrt{12}$ . However, it is not clear how one would combine such estimates together in a generalization of Expression 1 on page 111. The ISO/NIST guidance (ISO 1993; Taylor and Kuyatt 1994; JCGM 2006) does not address the problem of interval uncertainty—or any other kind of uncertainty—within the individual values  $X_i$  that are combined together using Expression 1. They are presumed to be point estimates. Nevertheless, within the spirit of the guidance, an analyst might construct a model of the measurement and its uncertainty as the distribution of means of samples of size  $N$  taken from a (equally weighted) mixture of uniform distributions representing the  $N$  respective intervals. This distribution could be derived analytically or via bootstrap methods or Monte Carlo simulation (JCGM 2006).

In the case of the skinny and puffy data sets, the mixture distributions are the same as those depicted in Figure 7 on page 33. The analog of the ISO/NIST model for uncertainty based on uniform distributions representing the intervals would be the distribution of means of random samples taken from these mixture distributions, where the sample size is 6 for skinny or 9 for puffy. No matter what the shape of the mixture distribution, the associated standard uncertainty can be estimated as the square root of its variance divided by the sample size. For the skinny data set, the standard uncertainty falls from 1.361 when incertitude is ignored to 1.244 when it is modeled with uniform distributions. In the case of the puffy data, the standard uncertainty grows from 0.6768 when incertitude is ignored to 0.7118 when it is modeled with uniform distributions. Note, however, that these changes are rather small. Whether the model of measurement uncertainty is constructed as a normal distribution with the mean of the midpoints and these standard uncertainties as the standard deviations, or a Monte Carlo simulation is



used to estimate the distributions of means of samples from the mixture distributions directly, the resulting distributions are graphically indistinguishable from the black ISO/NIST distributions shown in Figure 39 that ignore uncertainty altogether. Thus, we find that a conscientious analyst motivated to take account of the uncertainty in the two data sets would get answers very similar to those that ignore the uncertainty completely. In short, even when uncertainty is included, it seems not to have much of an impact on the results.

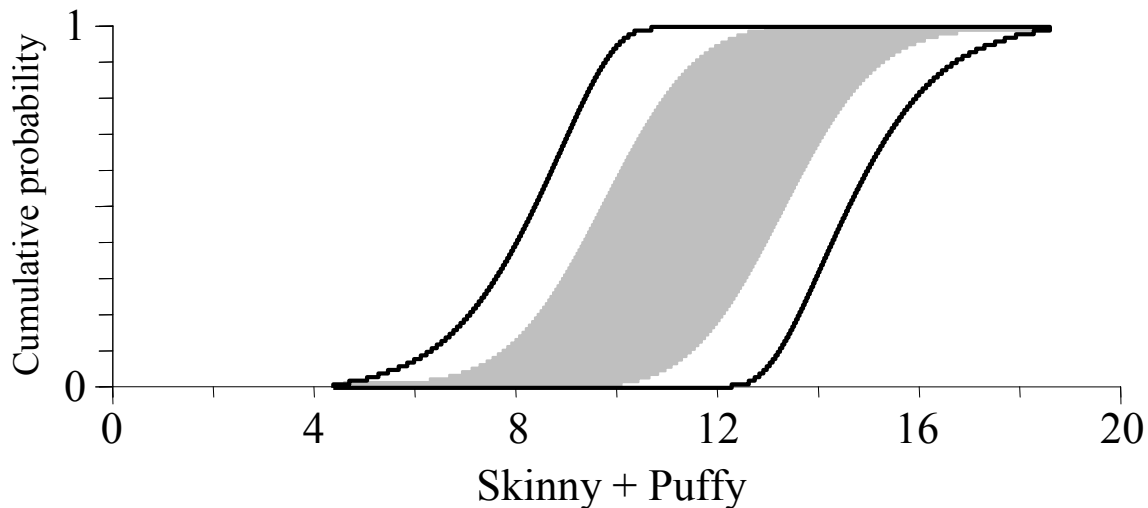
### 8.2.1 Dependencies among sources of uncertainty

The provisions within the ISO/NIST approach for handling possible correlations among uncertainties is presumably regarded as sufficiently general for wide use because it allows any correlations to be specified. Indeed, this flexibility is probably greater than many analysts are likely to ever need in practice. After all, observations from simultaneous measurements that could parameterize such correlations directly are rare. But we argue that this *parametric flexibility* is not really the same as, or as important as, *freedom from assumptions*. Although a user can specify correlation between sources of uncertainty, this option does not imply the user escapes having to make any assumptions about cross dependence among the sources of measurement uncertainty. It is telling that this flexibility often devolves in practice to choosing between only three correlations, zero and plus or minus one, corresponding to independence and perfect dependence. This may be because empiricists rarely have any more specific information that would point to another particular correlation coefficient inside the unit interval. Even in the face of empirical evidence on the subject of correlation, there seems to be an odd prejudice against modeling them among proponents of the standard methods. Dieck (1997, pages 148-151) says that calculated correlation coefficients should not be used unless they are statistically significantly different from zero. If there are few observations, even substantial correlations can fail to be statistically significant. The decision to set all non-significant correlations to zero means that some uncertainties will likely be underestimated and some will likely be overestimated.

JCGM (2006, section 9.4.1.3) acknowledges that engineers may sometimes have difficulty in quantifying the covariances or correlations. The guidance suggests treating inputs as independent when there is insufficient information to assess the correlation between them (ISO 1993, section F.1.2.1.c), but this is obviously incompatible with the recognition that correlations can have noticeable impacts on the results of calculations (JCGM 2006, section 9.4.3.3). We suggest that it might be useful to have a more appropriate method that could make the requisite computations when the analyst is unsure about what correlation or dependence there might be between inputs. There are two mathematical techniques for computing convolutions that do not require any assumptions about dependence between variables. The first is interval analysis (Wise and Henrion 1986), and the second is the probabilistic approach due to Fréchet (1935; 1951) of bounding results over all possible dependencies (Frank et al. 1987; Williamson and Downs 1990; Berleant and Goodman-Strauss 1998; Ferson et al. 2004a). Only these approaches truly relax dependence assumptions. The standard methods are making assumptions about the interactions of the sources of uncertainty that may or may not be supported by evidence and which may or may not be true in fact. Although it may not

often be necessary to make calculations without any assumptions about dependence between variables, it will be useful to be able to do so in some situations. Notable examples might include models of measurement devices functioning in abnormal environments or other cases where prior empirical information is limited or nonexistent.

*Numerical example.* Figure 41 depicts two results from propagating through an addition operation the uncertainties in the puffy and skinny data sets that were expressed via interval statistics as the gray p-boxes in Figure 39. Both were computed using algorithms described by Williamson and Downs (1990; Ferson et al. 2004a). The first is the result of addition assuming independence. This is shown with gray shading, and it is the same as was previously displayed in Figure 40. The second p-box, outlined in black, is the result of the same addition but *without any assumption at all about the dependence* between the two uncertainties. These outer bounds, therefore, simultaneously express the uncertainty arising from the incertitude of the input intervals, the within-data-set variability among them, and the compounded uncertainty of not knowing the correlation or dependence that interconnects the two estimates. Note that the uncertainty grows on account of the additional uncertainty about dependence, but it does not grow very much considering that the assumptions about dependence have been totally discharged. Inspecting the graphs, one can see that the interval ranging between the smallest possible 2.5<sup>th</sup> percentile and the largest possible 97.5<sup>th</sup> percentile is [6.5, 16.6] under the independence assumption. Without any assumption about dependence, this interval widens somewhat to [4.9, 18.1].



**Figure 41.** Uncertainty according to interval statistics about the sum of skinny and puffy assuming independence (gray) and making no assumption about dependence (black).

The standard approach (ISO 1993; Taylor and Kuyatt 1994) cannot account for uncertainty about the correlation or dependence among inputs. Nor, indeed, can any Monte Carlo method (JCGM 2006) account for this structural uncertainty. The outer bounds depicted in Figure 43 cannot be found by merely varying the correlation coefficient over its possible range on  $[-1, +1]$ ; such bounds would be too small and would neglect possible nonlinear dependencies between the inputs. Obtaining the correct bounds requires special-purpose (although straightforward) computation. This and analogous computations for subtraction, multiplication, division and other

mathematical operations can be made using the methods and algorithms spelled out in (Williamson and Downs 1990; Berleant and Goodman-Strauss 1998; Ferson et al. 2004a). If the analyst can assume that only linear correlations are possible between the uncertainties, or that at least the sign if not the magnitude of the dependence is known, then it is possible to tighten the bounds somewhat and reduce the overall breadth of the resulting uncertain number (Ferson et al. 2004a). The methods can also be used if specific information about the correlation between the inputs is available to obtain still tighter bounds, even if there is some incertitude about the inputs themselves. However, these methods are currently limited to cases in which the model combining the inputs is known explicitly, and would not be applicable if only partial derivatives of the function were known or if the function were implemented in a black box accessible only through Delphic query.

### **8.2.2 Accounting for incertitude in measurement uncertainty**

Because the standard ISO/NIST approach gives answers that are so clearly different from those obtained under interval statistics, we are lead immediately to several practical questions. Why are they different? How should we understand this difference, and should we think of using these two approaches for different purposes or in different circumstances? To answer these questions, we must consider the fundamental ideas behind the two approaches. A basic premise of interval statistics is the notion that the epistemic uncertainty (lack of perfect knowledge) embodied in incertitude of data intervals is fundamentally different from the aleatory uncertainty (variability) among the data values. Interval statistics is constructed to maintain this distinction and to treat the two forms of uncertainty differently. This report argues that incertitude (see especially section 2) cannot always be intuitively modeled as though it were the same thing as sampling uncertainty.

In contrast, the ISO/NIST guidance “treats uncertainty components arising from random effects and from corrections for systematic effects in exactly the same way” and, indeed, holds the viewpoint that “all components of uncertainty are of the same nature and are to be treated identically” (ISO 1993, section E.3). This viewpoint is based on the subjectivist interpretation that sees probability as a degree of belief rather than anything necessarily realized in the physical world. It was adopted in reaction to the limitations of a purely frequentist approach that could not handle non-statistical uncertainties (ibid.). The flexibility of the ISO/NIST approach based on the subjectivist Bayesian interpretation of probability depends on the presumption that the only kind of uncertainty that exists in a measurement is equivalent to the kind that can be properly characterized as sampling uncertainty among repeated measurings. Yet one doesn’t need to conflate the two kinds of uncertainty to gain the desired flexibility. Interval statistics allows an analyst to distinguish between them when doing so may be useful.

We need not take sides in the long-standing philosophical debate between frequentists and subjectivist Bayesians. The question of what to do about incertitude may be orthogonal to this debate, at least in part. We feel that interval statistics offers methods that can be useful under *either* interpretation of probability. In principle, many of the

ideas in this report are compatible with both frequentist and subjectivist Bayesian\* approaches (except, for instance, the purely frequentist notions such as Neyman-Pearson confidence intervals). In fact, both frequentist and Bayesian approaches could use interval statistics to improve the job they do of expressing uncertainty in situations where both aleatory uncertainty (variability) and epistemic uncertainty (lack of knowledge) are present in measurements.

As we and many others have argued, no single probability distribution can properly represent the epistemic uncertainty characterized by an interval (Ahmad 1975; Shafer 1976; Walley 1991; Ferson and Ginzburg 1996; Horowitz and Manski 2000; Manski 2003; Ferson et al. 2003; Ferson and Hajagos 2004; Kreinovich 2004; Kreinovich and Longpré 2004; Starks et al. 2004; Hajagos 2005, *inter alia*). The examples given in section 8.2 and elsewhere in this report suggest that ignoring incertitude, or modeling it with uniform distributions, can produce numerical results that violate our intuition about just what it means to be ignorant about something. How large the incertitude within a data set is should make a difference in the inferences that can be drawn from the data. Although it may often be the case that one can ignore the incertitude, this is surely not always true. Therefore, we suggest that methods that ignore incertitude or assume uniformity over it must be viewed as approximations when they produce results that disagree with those of an interval statistics approach, which is the more comprehensive analysis. Such comprehensiveness comes at a price; obtaining results in interval statistics can be more expensive computationally than are approximations.

We emphasize that we are not simply calling for empiricists to recognize larger uncertainties than they usually acknowledge. Indeed, one's uncertainty may be either small or large. Instead, we are merely arguing that a system by which such uncertainties are represented, combined and propagated might benefit from generalization to account for various *kinds* of uncertainties that currently are being ignored. Chiefly of concern is the incertitude that comes from intermittent or finite precision observations, censoring, security or privacy constraints, etc., such as were described in section 2. We certainly feel that analysts should be free to make whatever assumptions they like. Our main concern is that they should not be forced to make assumptions they are not comfortable with, merely to be able to compute an answer. Interval statistics can help to relax the assumption is that resolution uncertainty for individual measurements is negligible compared to the variability among repeated measurements. This assumption would seem to limit the application of the ISO/NIST approach to a narrow swath of mature disciplines within science and engineering, and preclude its use in cutting-edge domains where errors have not yet been brought under good experimental control. The primary advantage of an interval statistics approach relative to the ISO/NIST standard approach is that it relaxes the assumptions that these measurement uncertainties are relatively small.

---

\*It should be emphasized that handling incertitude differently is not necessarily antithetical to a Bayesian approach. The subjectivist definition of probability as a measure of the degree of belief that an event will occur does not preclude being agnostic about one's belief. Recent work in robust Bayes analysis (Insua and Ruggeri 2000; Pericchi 2000; Berger 1994; Pericchi and Walley 1991; Berger 1985) demonstrates that Bayesian methods can be usefully combined with methods that treat interval-like uncertainty.

### 8.3 Harmonization

We have argued that the methods of interval statistics reviewed in this document constitute an *alternative* to the current practices for representing and propagating measurement uncertainty that are recommended by national and international standards bodies. This section will discuss a slightly deeper *harmonization* of these two approaches. This harmonization will integrate concepts from both interval statistics and the ISO/NIST approach to modeling measurement uncertainty and might constitute a hybrid approach for representing and propagating measurement uncertainty that has the advantages of both approaches in allowing for substantial interval uncertainty and yet also acknowledging that errors can be very large or even unbounded. The key to harmonizing the two approaches that will unlock this hybrid vigor is to generalize the model of a measurement's possible error. The ISO/NIST approach uses a normal distribution, and the interval statistics approach uses an interval. A better general model for measurement uncertainty would have features of both a probability distribution and an interval. Estler (1997) suggested a distribution-free model for measurement uncertainty that has these features.

Figure 42 shows a variety of possible models of the uncertainty associated with a measurement or even a single datum. The upper, left graph is a precise probability distribution, which is the traditional model for uncertainty in the ISO/NIST approach (where the standard deviation is the observed standard deviation of a population of similar measurements). A symmetric normal shape is depicted, but other distribution shapes are also possible, uniforms and triangulars being especially common. It seems clear that this model of measurement uncertainty will continue to be important in a host of practical settings even if interval statistics is embraced by the metrological community. The upper, right graph of the figure shows an interval such has been used throughout this report to model a particular kind of measurement uncertainty which we have called *incertitude*. We think the argument is compelling for a central role for intervals in models for measurement uncertainty in situations where *incertitude* cannot be neglected.

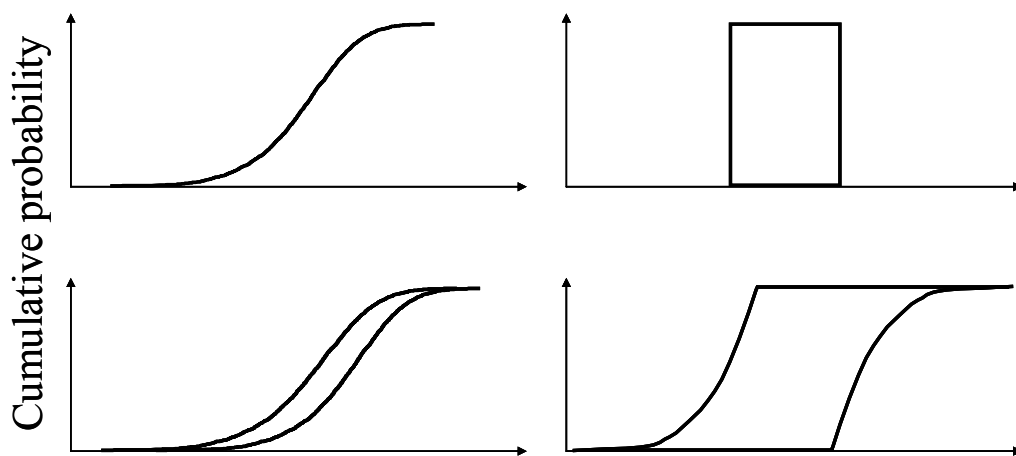


Figure 42. Possible models for measurement uncertainty about a single measurement (x-axes), including the ISO/NIST normal distribution (upper left), the interval of interval statistics (upper right), and two hybrids (lower left and right).

The two lower graphs of the figure depict possible generalizations of the precise probability distribution and the interval. These objects represent uncertainty about the precise location of the probability distribution of error. Mathematically, they are p-boxes, and they can also be characterized as Dempster-Shafer structures (Ferson et al. 2003). We have seen several examples of the structure in the lower, left graph of the figure in section 8.2. This structure can arise whenever quantities characterized by a probability distribution and an interval are combined in convolutions. Thus, such objects are quite natural in any problem that contains both randomness and incertitude. Note, however, if the left and right bounds are normal distributions, this structure does not generalize an interval because, no matter how far apart the bounds are, there is no core, i.e., no region of values that are associated with vacuous limits on probability. On the other hand, the lower, right graph, which might be thought of as an interval with tails, does have a core region of interval uncertainty representing empirical incertitude. Such an object commonly arises from enveloping probability distributions with intervals (rather than convolving them). As important as having the core of epistemic uncertainty, this structure's tails also acknowledge that errors are not *guaranteed* to be within this narrow range. This, finally, is the promised relaxation of the full confidence assumption we introduced on page 18. It might be an appropriate model of measurement uncertainty whenever an empiricist acknowledges significant incertitude in a measurement but is also aware that a measurement might be really quite wrong because of the possibility of blunders, transcription errors, system noise, and other sources. The tails allow the analyst to account for his or her lack of 'full' confidence about a reported interval. The objects in the lower graphs are not otherwise qualitatively different, nor do they require different mathematical methods to handle them. They are two instances of a general class of objects that capture both randomness as described by probability and incertitude as described by interval analysis.

Any of the models in the figure could be used to represent the measurement imprecision of an individual quantification. More importantly, collections of these models can be combined together and their uncertainty propagated through calculations. Moreover, different calculi are not required to handle distributions, intervals and the more general hybrids; they can all be treated with a single computational technology. Our previous report (Ferson et al. 2003) introduces these objects and explains how they can be used in mathematical calculations. It is relatively straightforward to form distributions of these models, which generalize the distributions of intervals considered in section 4.1, which themselves generalized empirical distribution functions based on scalar values. Recent work has also shown that elementary descriptive statistics for collections of these objects can be computed efficiently (Kreinovich et al. 2006b). It seems likely that the use of these generalized models of measurement uncertainty can even be extended into use with inferential statistical methods, although this will require research to demonstrate conclusively.

The operational details of how a fully hybridized system would work in routine practice remain to be explored. It seems reasonable that canonical models of uncertainty might be selected to recommend for routine use. For instance, in circumstances where the normal

distribution is now recommended by the ISO/NIST methodology, a p-box qualitatively like that in the lower, left graph of Figure 42 might be recommended instead. What is needed to make the hybridization into a successful marriage in practice is the development of *conventions and traditions* in thinking about what appropriate uses will be for the different possible models of uncertainty. Clearly, such development will emerge from a process involving many researchers, but the first step is to recognize a need.

## **8.4 Is measurement uncertainty underestimated?**

This section entertains the question of whether there is any empirical evidence to think that the standard methods metrologists have devised for expressing measurement uncertainty may be missing something.

By arguing that incertitude is sometimes present in real-world measurements and that the standard methods for handling uncertainty are not properly accounting for incertitude, this report is suggesting that the expressed uncertainty of measurements may be underestimated in some number of cases. Is there any *evidence* that uncertainty is being underestimated by the standard methods currently used in metrology? It has long been recognized, as a matter of fact, that there seems to be. Shewart (1939) documented overconfidence in the uncertainty statements for estimates of the speed of light. Apparently this overconfidence has extended through the entire history of this measurement. Figure 43 depicts the uncertainty about historical measured values of the speed of light in a vacuum for various determinations made through time, as expressed by the measurement teams themselves. The precise meanings of the specified intervals were not always clear in the published reports, but, giving them the benefit of the doubt, we might take them to represent plus or minus one standard deviation (rather than 95% confidence intervals as is the standard format today). The nonlinear square-root scale of the velocity axis is centered on the current definition of the speed of light so that the individual ranges can more easily be distinguished as they progressively tighten through time. This scale obscures the tightening over the years as professed by the different measurement teams. They were getting more and more confident, though not necessarily more accurate in a relative sense. Henrion and Fischhoff (1986) revisited the issue of overconfidence and noted that, if these measurements were well calibrated, one would expect about 70% of them would enclose the true value. Instead, fewer than half (13 out of 27) of these ranges include the currently accepted\* value of the speed of light.

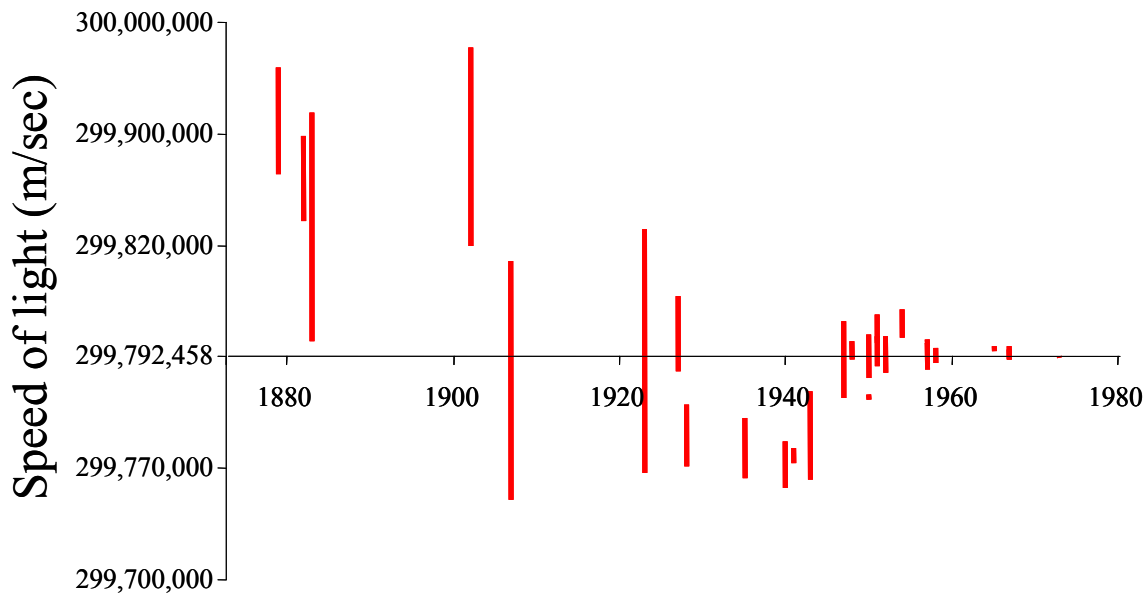
Youden (1972; Stigler 1996) similarly described overconfidence—which is to say, systematic underestimation of uncertainty—in the measurements of both the speed of light and the astronomical unit (mean distance from the earth to the sun). Youden's famous address is widely cited in many scientific disciplines and has sparked lots of commentary, some of which may be overstated on account of confusion about the

---

\*In 1983, the uncertainty about the speed of light officially became zero. That year the SI unit meter was redefined in terms of  $c$ , the speed of light in a vacuum, rendering the speed of light to be a defined value at exactly 299,792,458 meters per second. Therefore the speed of light is now a constant by definition without any empirical uncertainty. How convenient!

meaning of confidence intervals (Killeen 2005). Henrion and Fischhoff (1986) also examined the “performance” of physicists in assessing uncertainty due to possible systematic errors in measurements of physical quantities by comparing historical measurements against the currently accepted values (i.e., values recommended by standards agencies) for a variety of fundamental physical constants. They observed consistent underestimation of uncertainties. Morgan and Henrion (1990, page 59) asserted that the overconfidence “has been found to be almost universal in all measurements of physical quantities that have been looked at.” This overconfidence is apparently pervasive in science generally and may be related to human abilities systemically. It is well known in the psychometric literature that lay people and even experts are routinely and strongly overconfident about their estimates (Plous 1993). When experts are asked to provide 90% confidence intervals for their judgments (which ought to enclose the true value 90% of the time on average), their ranges will enclose the truth typically only about 30 to 50% of the time. Shlyakhter (1993; 1994) suggested that this overconfidence is so pervasive in science that we might introduce an automatic inflation factor to all uncertainty statements to account for it.

Of course this brief discussion cannot pretend to be compelling on the question about whether uncertainties are correctly estimated in practice, but it hopes to suggest, against a mountain of comfort with the status quo, that the question is legitimate and worth considering. If substantial miscalculation of uncertainties is routinely or even regularly occurring, it is inconceivable that it wouldn’t matter to metrologists and engineers.



**Figure 43. Measurement uncertainty for the speed of light at different years. Note the nonlinear (square root) scale of the ordinate. Redrawn from Cutler (2001a).**



## 9 Conclusions and next steps

In a variety of common circumstances, interval uncertainty (incertitude) arises in real-world measurements. The mensurational and statistical phenomena of censoring, rounding, data binning, and intermittency or periodicity of measurements described in section 2 strongly imply that intervals, rather than any probability distributions, are the proper mathematical representations of the uncertainties of some measurements in many practical cases. Using intervals to express and propagate these kinds of uncertainties is a relatively inexpensive way to make a formal accounting of the impact of measurement uncertainty. In interval statistics, then, both sampling uncertainty and measurement uncertainty can be expressed in a way that distinguishes the two forms. We argue that incertitude should be incorporated into statistical analyses whenever it is non-negligibly large for the following three reasons:

- (1) Incertitude can affect statistical conclusions based on these measurements.
- (2) It is inappropriate to mix together data of different quality (narrow and wide intervals) without accounting for their differences.
- (3) Incertitude can interact with variability (scatter) of the data in complicated ways that result in tradeoffs between measurement precision and sample size.

Some statistics for interval data are straightforward to compute. Others can often be cumbersome to compute, and some are computationally very difficult. Section 6 summarizes the computability of some common statistics for various kinds of data sets containing incertitude. These considerations become important when the data become abundant.

***Alternative measurement theory.*** Interval statistics provides an alternative to the purely probabilistic treatment of measurement uncertainty. Sometimes it is necessary to combine highly precise measurement results with poor precision measurements in calculations. Practically minded analysts often worry whether it is reasonable to mingle together poor data with good data. They wonder how they should understand the quality of the computed results. We hold that it is wrong to mix together good and poor data without accounting for their different reliabilities. Doing so would mask the uncertainty of the poor data and dilute the informativeness of the good data. Traditional approaches might employ weighting schemes to differentiate between good and poor data. Interval statistics permits this disparate information to be combined in a coherent way that makes the consequent uncertainty of the result completely transparent in the width of its interval. The difference between interval and traditional statistics becomes stark if we consider the behavior of uncertainty estimates as the number of samples grows asymptotically large. In the traditional probabilistic approach, the uncertainty of the mean, for instance, becomes vanishingly small, even if the individual measurements have a great deal of uncertainty. In interval statistics, on the other hand, the incertitude from plus-minus ranges limit the precision with which calculations of statistics can be made *no matter how large the sample size is*. If I can only measure a quantity to plus or minus 1000 units, my estimate of the uncertainty about the mean cannot really get much better than 1000 units. Historians of science have long recognized that expressions of uncertainty for many empirical measurements are overconfident. The tradition of ignoring incertitude may be

part of the reason for this. The standard approach to modeling measurement uncertainty is not quite yet written in stone. NIST policy, for instance, holds that “It is understood that any valid statistical method that is technically justified under the existing circumstances may be used to determine the equivalent of [quantitative measures of uncertainty]” (Taylor and Kuyatt 1994, Appendix C, section 4). Recent supplementary guidance (JCGM 2006) on using Monte Carlo methods illustrates, if nothing else, that metrologists are interested in generalizing their methods when it is useful to do so.

***Alternative for censored and missing data methods.*** Interval statistics also provides an alternative treatment for data censoring and missing values. In traditional probabilistic approaches, missing or censored values are often presumed to have the same distribution as the available data. Although this is not refutable by appeal to the data themselves, the assumption may still not be credible (Manski 2003; Zaffalon 2002; 2005). Analysts find it hard—and perhaps *should* find it hard—to justify this assumption, or any similar assumption that identifies a particular distribution for missing or censored data. The reason the data are missing in the first place may be related to their magnitudes. But an even more fundamental consideration ought to restrain the analyst’s assumptions. It is the impetus behind science itself, to wit, that *observations* rather than guesses should guide inferences. An epistemology that allows us to make any assertion in the absence of information, in the end, doesn’t compel us to believe any of its conclusions.

***Experimental design.*** The work should also have implications for experimental design. It is often true that a coarse measurement can be made cheaply whereas a careful measurement is more expensive. This suggests there should be a tradeoff between sample size and precision within an empirical effort. Yet many practicing statisticians seem to think the tradeoff always favors increasing sample size. We argue that this is clearly wrong. The tradeoffs may be nonlinear, so that whether an incremental addition to the budget for some empirical effort should be allocated to increasing sample size or improving precision will depend on the quantitative details of the problem and, in particular, what the current budget is. Sometimes the optimal strategy will increase the number of samples to be collected, and sometimes it will increase the care devoted to the individual measurements, and sometimes it might do some of both. The methods of interval statistics should allow analysts to design empirical efforts to optimize such allocations, in effect choosing between more data or better data, to maximize the benefit on the overall conclusion. If this tradeoff is not explored as a part of experimental design, it is likely that either sampling effort or precision will be wasted. Common statistical methods such as power analyses focus on the sampling uncertainty but tend to neglect measurement uncertainty, and may even ignore uncertainty all together.

***Distinguishing aleatory and epistemic uncertainty.*** In modern risk assessments and uncertainty modeling generally, analysts conscientiously distinguish between epistemic and aleatory uncertainty. This conscientiousness about the distinction should extend to data handling and statistical methods as well, but this has generally not been the case. For instance, guidance promulgated by national and international standards agencies for handling uncertainty in empirical data does not adequately distinguish between the two forms of uncertainty and tends to confound them together. Likewise, because most of the

statistical methods developed over the last century have focused on assessing and projecting sampling uncertainty (which arises from having observed only a subset of a population) and neglect measurement uncertainty (which arises from imperfect mensuration, censoring or missing values), almost all of the statistical methods widely employed in risk analysis presume that individual measurements are infinitely precise, even though this is rarely justifiable.

## **9.1 *What should we call this?***

It might be worth reflecting on what the subject matter of this report should be called. We might simply say that this material is a part of the field of robust statistics, but this masks the unity and specific uses of the present approach. The phrase “imprecise statistics” is evocative and would denote the field of statistics of imprecise data, just as circular statistics is the sister field focusing on circular (periodic) data. However, this name seems doomed to produce misunderstanding because the phrase can also be parsed to mean “poor or ill-posed statistics”. In the recent past, we have seen the phrase “fuzzy logic” devolve from denoting the mathematical methods that can handle fuzzy specifications in an exact way into misapprehension as referring to clouded or invalid logic, probably because of the cognate phrase “fuzzy thinking”.

Manski’s (2003) phrase “statistics of incomplete data” seems too focused narrowly on missing data problems. Because the general issue concerns identifying distributions and their statistics insofar as possible based on only partial information, Manski also uses the phrase “partial identification” to characterize the basic problem, but this sounds perhaps too much like a riddle rather than an informative name.

In this report, we have settled on the phrase “interval statistics” even though it is not perfect. Firstly, it does not capture the generalization to using p-boxes to represent measurement uncertainty. What’s worse is that the name “interval statistics” runs the risk of being misunderstood as somehow applying only to intervals and not being relevant to studies involving variability or probabilistic ideas. We have noticed in the past that prejudice against interval analysis exists in some quantitative disciplines and simply using the word “interval” sometimes engenders distrust or rejection of the method. The methods and ideas discussed in this report are, of course, intended to bridge interval analysis and traditional statistics based exclusively on probability theory. Perhaps in future developments the phrase “statistics of imprecise data” will be preferred.

## **9.2 *Next steps***

Research effort is needed to derive practical algorithms for computing interval generalizations of some remaining important statistics and statistical tests. Most useful immediately would be interval versions of analysis of variance and *t*-tests for both unpaired and paired data. Development effort is also needed to implement these algorithms, including those described in this report, in a friendly statistical software package so that they will be accessible to engineers widely.

Methodological research is needed on the integration of interval statistics into the standard protocols for assessing and propagating measurement uncertainty so that this complementary approach can be deployed in uncertainty-critical situations. It will be important to develop case studies in which the competing approaches are compared, side by side, so that the advantages of each method are highlighted and the prospects for synthesis between them are clearly illuminated. It will also be important to explore exactly how a hybridization of interval statistics with the standard ISO/NIST approach could work in practice. In generalizing the normal distribution of the ISO/NIST approach to account for epistemic uncertainty like incertitude, we might elect to use a p-box or Dempster-Shafer structure on the real line as the model of knowledge and uncertainty for an individual datum, i.e., for each measurement. This would generalize both the normal distributions used in the standard approach and the intervals used in this report. Kreinovich et al. (2006b) have discussed the prospects of computing descriptive statistics on such objects. Further research is needed to address several questions before such a hybridization could be recommended for routine use in metrology. For instance, would a simple generalization of a normal distribution that has interval parameters be workably flexible model of uncertainty in situations where incertitude is non-negligible in measurements? Is there a need to represent intervals as distinct from such generalizations of normal distributions? Should the law of uncertainty propagation be considered a definition or merely an approximation? If the latter, what practical criteria should govern whether it or a more accurate formula is used to compute the uncertainty of derived variables? Can calculations that make no assumptions about dependence between input variables lead to usefully tight characterizations of uncertainty for derived variables? What models and constructs beyond those already in ISO/NIST guidance could or should be recommended for routine use in measurements? The approach currently recommended is the product of decades of use and refinement. A graceful generalization that could achieve the goal of a more comprehensive representation of incertitude as well as other forms of aleatory and epistemic uncertainty may require attention to other details and consideration of further subtleties, and may therefore need some time to mature.

## 10 Appendix

The following is the text of Recommendation INC-1 (1980) of the Working Group on the Statement of Uncertainties convened by International Bureau of Weights and Measures (BIPM, Bureau International des Poids et Mesures) at the behest of the International Committee for Weights and Measures (CIPM, Comité International des Poids et Mesures) and approved by it, as captured from the National Institute of Standards and Technology website <http://physics.nist.gov/cuu/Uncertainty/international1.html>. See Kaarls (1981).

### **Recommendation INC-1 (1980)**

#### Expression of experimental uncertainties

1. The uncertainty in the result of a measurement generally consists of several components which may be grouped into two categories according to the way in which their numerical value is estimated.

Type A. Those which are evaluated by statistical methods

Type B. Those which are evaluated by other means

There is not always a simple correspondence between the classification into categories A or B and the previously used classification into “random” and “systematic” uncertainties. The term “systematic uncertainty” can be misleading and should be avoided. Any detailed report of uncertainty should consist of a complete list of the components, specifying for each the method used to obtain its numerical value.

2. The components in category A are characterized by the estimated variances  $s_i^2$  ( or the estimated “standard deviations”  $s_i$ ) and the number of degrees of freedom  $\nu_i$ . Where appropriate the covariances should be given.

3. The components in category B are characterized by quantities  $u_j^2$ , which are considered to be approximations to the corresponding variances, the existence of which is assumed. The quantities  $u_j^2$  may be treated like variances and the quantities  $u_j$  like standard deviations. Where appropriate, the covariances should be treated in a similar way.

4. The combined uncertainty should be characterized by the numerical value obtained by applying the usual method for the combination of variances. The combined uncertainty and its components should be expressed in the form of “standard deviations.”

5. If for particular applications, it is necessary to multiply the combined uncertainty by an overall uncertainty, the multiplying factor must always be stated.

## 11 Glossary

**accuracy** The closeness of a measurement to its measurand's true value.

**aleatory uncertainty** The kind of uncertainty resulting from randomness or unpredictability due to stochasticity. Aleatory uncertainty is also known as variability, stochastic uncertainty, Type I or Type A uncertainty, irreducible uncertainty, conflict, and objective uncertainty

**best possible** As tight as can be justified (said of bounds). An upper bound is best possible if it is the smallest such bound possible, and a lower bound is best possible if it is the largest lower bound possible.

**censoring** Imprecision in data such that only upper bounds are known (left censored data or non-detects), lower bounds are known (right censored or suspended data), or non-equivalent upper and lower bounds are known (interval censored data).

**CIPM** International Committee for Weights and Measures, which is the highest authority in the field of metrology.

**cumulative distribution function** A distribution function.

**D-parameter** A real-valued statistic or function  $D$  of a distribution function or a random variable such that  $D(F) \leq D(G)$  whenever  $F$  stochastically dominates  $G$ . A D-parameter is said to respect stochastic dominance.

**data** (singular: datum) Empirical information, usually in numerical form. In this report, data may be scalar (point) values or intervals, including the vacuous intervals  $[-\infty, \infty]$  and  $[0, \infty]$ .

**Dempster-Shafer structure** A set of focal elements (in this report, closed intervals of the real line), each of which is associated with a non-negative real values that sum to unity.

**Dempster-Shafer theory** A variant of probability theory in which the elements of the probability space to which nonzero mass is attributed, called focal elements, are not singletons but rather sets which represent the indistinguishability of alternatives within bodies of evidence. Dempster-Shafer theory is sometimes called evidence theory.

**distribution function** The function  $F$  associated with some variate that describes the probability  $F(X)$  that the variate will take on a value not greater than  $X$ . The distribution function associated with a data set of scalar values describes the probability  $F(X)$  that a value selected at random (i.e., uniformly and independently)

from the data values will have a value less than or equal to  $X$ . Also known as a cumulative distribution function.

**dominates** Stochastically dominates.

**EDF** Empirical distribution function.

**empirical distribution function** A distribution function for a data set. Sometimes abbreviated EDF.

**epistemic uncertainty** The kind of uncertainty arising from imperfect knowledge, rather than variability. Also called ignorance, subjective uncertainty, Type II or Type B uncertainty, reducible uncertainty and state-of-knowledge uncertainty.

**error** The difference between a measurement and its measurand's true value.

**Gaussian distribution** Normal distribution.

**incertitude** Uncertainty characterized by an interval (rather than a probability distribution).

**interval** The set of all real numbers lying between two fixed numbers called the endpoints of the interval. In this report, intervals are usually closed so that the endpoints are considered part of the set. An interval is sometimes conceived as a contiguous subset of the extended real numbers, so that its endpoints may include  $-\infty$  or  $+\infty$ . An alternative definition of an interval  $[A, B]$  is the set of all cumulative probability distribution functions  $F$  defined on the real numbers such that  $F(X) = 0$  for all values  $X < A$ , and  $F(X) = 1$  for all values  $X > B$ . These alternative definitions lead to distinct theories about intervals.

**interval data** In this report, data that may include values that are intervals.

**ISO** International Organization for Standardization, a non-governmental organization of national standards institutes headquartered in Geneva.

**lexicographic order** Ordering of elements in a Cartesian product space such that the element  $A$  is less than or equal to another element  $B$  whenever the first component of  $A$  is less than or equal to the first component of  $B$ , or the first  $k$  components of  $A$  and  $B$  are equal and the  $k+1^{\text{st}}$  component of  $A$  is less than or equal to the  $k+1^{\text{st}}$  component of  $B$ . Also known as dictionary or alphabetic order because it is how dictionaries arrange words (which are elements of a Cartesian product space of letters).

**mean** The probability-weighted average of a set of values or a probability distribution. The mean is also called the expected value or the expectation of a random variable. It is the first moment of a probability distribution.

**measurand** A specific quantity subjected to measurement.

**measurement** The value resulting from an approximation or estimation of a quantity, or, sometimes, the process of approximating or estimating a quantity.

**measurement error** The difference between a measured quantity and its actual or true value. The phrase is also sometimes used to refer to the imprecision or uncertainty about a measurement, although the phrase measurement uncertainty is now preferable for this meaning.

**measurement uncertainty** The uncertainty about the accuracy of a measurement.

**median** A magnitude that is greater than half of the values, and less than half of the values, in a data set or taken on by a random number. This value is the 50<sup>th</sup> percentile and the 0.5 quantile of a probability distribution.

**mensuration** The act or process of taking a measurement.

**metrology** The science of measurement.

**missing data** Unobserved or lost values in a data set. A missing value can be represented by a vacuous interval  $[-\infty, \infty]$ , or  $[0, \infty]$  if the variable is surely nonnegative.

**mode** Value that occurs most frequently in a set of values or a probability distribution.

**naive interval analysis** The evaluation of mathematical expressions by sequential application of the elementary rules of interval arithmetic, without regard to the possibility that the result can become unnecessarily wide in some situations.

**narrowed intervals** Given any interval data set  $x_i, i = 1, \dots, N$ , the transformed intervals

$$y_i = \left[ \tilde{x}_i - \gamma \frac{\Delta_i}{N}, \tilde{x}_i + \gamma \frac{\Delta_i}{N} \right] = \left[ \underline{x}_i + \bar{x}_i - \gamma \frac{(\bar{x}_i - \underline{x}_i)}{N}, \underline{x}_i + \bar{x}_i + \gamma \frac{(\bar{x}_i - \underline{x}_i)}{N} \right] / 2,$$

where  $i = 1, \dots, N$ ,  $\gamma$  is a constant, and  $\tilde{x} = (\underline{x} + \bar{x})/2$  is an interval's midpoint, and  $\Delta = (\bar{x} - \underline{x})/2$  is its halfwidth. The narrowed intervals  $y_i$  have the same respective midpoints as the original data  $x_i$ , but they have smaller widths, depending on  $\gamma$  and the sample size  $N$ .

**NIST** National Institute of Standards and Technology, an agency of the United States Department of Commerce.

**non-detect** An observation below the limit of detection or limit of resolution of a measurement device or protocol.



**normal distribution** A probability distribution having a particular shape from a family of distributions characterized by continuous, unimodal and symmetric density with infinite tails in both directions and parameterized by a mean and standard deviation.

**NP-hard** At least as difficult computationally as any problem in the NP (nondeterministic polynomial time) class of problems. It is widely presumed that every algorithm that solves all cases of an NP-hard problem requires exponential computational time, growing in complexity with sample size or other problem feature.

**outward-directed rounding** The convention used to preserve rigorousness of interval calculations and outputs by rounding the lower endpoints of intervals downward and the upper endpoints of intervals upward (rather than in each case to the nearest number in some number of digits).

**optimal** Best possible.

**p-box** A probability box.

**percentile** A value that divides the range of a set of data or a distribution such that a specified percentage of the data or distribution lies below this value.

**precision** The narrowness of uncertainty about, or the reproducibility of, a measurement.

**probability bounds analysis** An analysis or calculation involving interval probabilities or probability boxes.

**probability box** A class of distribution functions  $F(X)$  specified by a bounding pair of distribution functions  $\underline{F}(X)$  and  $\bar{F}(X)$  such that  $\underline{F}(X) \leq F(X) \leq \bar{F}(X)$  for all  $X$  values.

**quantile** A value that divides the range of a set of data or a distribution such that a specified fraction of the data or distribution lies below this value.

**random sample** A set of values drawn from a population in such a way that, each time an item is drawn, every item in the population has an equal chance of being selected.

**random error** Error that occurs when recording a true value  $X$  as  $X+\varepsilon$  where  $\varepsilon$  is a realization of a random variable. When the mean of the random variable  $\varepsilon$  is zero, the random error is said to be unbiased. Contrasted with systematic error.

**real number** An element of the real line consisting of positive and negative integers, rational numbers, irrationals and transcendental numbers. A rational number or the limit of a sequence of rational numbers. Also called scalars.

**rigorous** Strict or sure, as opposed to merely approximate. Often said of bounds which can be rigorous without being best possible.

**sampling uncertainty** The uncertainty about a statistic or a probability distribution arising from incomplete sampling of the population being characterized by the statistic or distribution.

**scalar** A precise real number, or, sometimes, a precise rational number.

**standard deviation** A measure of the dispersion among data values computed as the square root of their variance.

**standard error** The standard deviation of a statistic. In this report, it refers to the standard deviation of the mean. Thus, it is a characterization of the standard deviation of a population of means (of some specified sample size) drawn at random from an underlying distribution of values.

**standard uncertainty** The estimated standard deviation associated with a measurement result. Usually, the measurement result is an average so the standard uncertainty is the standard deviation of a mean, which is also called the standard error.

**stochastically dominates** Having a cumulative distribution function entirely to the left of another cumulative distribution function (which is said to be dominated). If  $F$  and  $G$  are cumulative distribution functions, then  $F$  stochastically dominates  $G$  if  $F(X) \geq G(X)$  for all values of  $X$ .

**systematic error** The difference between the true value of a quantity and the value to which the mean of repeated measurements converges as more measurements are taken. It is the error that occurs when the result of measuring a quantity whose actual value is  $X$  appears to be  $f(X)$ , where  $f$  is a fixed function. Contrasted with random error.

**uncertain number** A numerical quantity or distribution about which there is uncertainty. Uncertain numbers include intervals, probability distributions, probability boxes, Dempster-Shafer structures as special cases. Uncertain numbers also include scalars (real numbers) as degenerate special cases.

**uncertainty** The absence of perfectly detailed knowledge. Uncertainty includes incertitude (the exact value is not known) and variability (the value is changing). Uncertainty may also include other forms such as vagueness, ambiguity and fuzziness (in the sense of border-line cases).

**variability** The fluctuation or variation due to randomness or stochasticity. Variability is also called aleatory uncertainty, stochastic uncertainty, Type 1 or Type A uncertainty, irreducible uncertainty and objective uncertainty.

**variance** A measure of the dispersion or variability of a set of data  $X_i, i = 1, \dots, N$ , as the average of squared deviations of the data from their arithmetic mean. When the data set effectively covers all the members of a population, the variance is computed as  $\Sigma(X_i - \Sigma X_i / N)^2 / N$ . This is called the population variance. When the data set is only a subset of the population considered, the variance is better estimated as  $\Sigma(X_i - \Sigma X_i / N)^2 / (N - 1)$ . This formulation is called the sample variance.

## 12 References

- Ahmad, R. 1975. A distribution-free interval mathematical analysis of probability density functions. Pages 127-134 in *Interval Mathematics*, G. Goos and J. Hartmanis (eds.). Lecture Notes in Computer Science, Springer-Verlag, Berlin.
- Akritis, M.G., S.A. Murphy and M.P. La Valley. 1995. The Theil-Sen estimator with doubly censored data and applications to astronomy. *Journal of the American Statistical Association* 90: 170-177.
- Alefeld, G., and J. Hertzberger. 1983. *Introduction to Interval Computations*. Academic Press, New York.
- Bandemer, H., and W. Näther. 1992. *Fuzzy Data Analysis*. Kluwer Academic Publishers, Dordrecht.
- Beck, J., V. Kreinovich and B. Wu. 2004. Interval-valued and fuzzy-valued random variables: from computing sample variances to computing sample covariances. Pages 85-92 in *Soft Methodology and Random Information Systems*, M. Lopez, M. A. Gil, P. Grzegorzewski, O. Hryniewicz, and J. Lawry (eds.), Springer.
- Beer, M. 2006. Sampling without probabilistic model. Pages 369-390 in *Proceedings of the NSF Workshop on Reliable Engineering Computing: Modeling Errors and Uncertainty in Engineering Computations*, R.L. Muhanna and R.L. Mullen (eds.), Savannah, Georgia. [http://www.gtrep.gatech.edu/workshop/rec06/papers/Beer\\_paper.pdf](http://www.gtrep.gatech.edu/workshop/rec06/papers/Beer_paper.pdf)
- Berger, J.O. 1985. *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New York.
- Berger, J.O. 1994. An overview of robust Bayesian analysis [with discussion]. *Test* 3: 5-124. <http://www.stat.duke.edu/~berger/papers/overview.ps>
- Berleant, D. 1993. Automatically verified reasoning with both intervals and probability density functions. *Interval Computations* 1993 (2): 48-70.
- Berleant, D. 1996. Automatically verified arithmetic on probability distributions and intervals. Pages 227-244 in *Applications of Interval Computations*, B. Kearfott and V. Kreinovich (eds.), Kluwer Academic Publishers.
- Berleant, D., and H. Cheng. 1998. A software tool for automatically verified operations on intervals and probability distributions. *Reliable Computing* 4: 71-82.
- Berleant, D., and C. Goodman-Strauss. 1998. Bounding the results of arithmetic operations on random variables of unknown dependency using intervals. *Reliable Computing* 4: 147-165.
- Berleant, D., L. Xie and J. Zhang. 2003. Statool: a tool for distribution envelope determination (DEnv), an interval-based algorithm for arithmetic on random variables. *Reliable Computing* 9(2): 91-108.
- Berleant, D., M. Ceberio, G. Xiang and V. Kreinovich. 2007. Towards adding probabilities and correlations to interval computations. *International Journal of Approximate Reasoning* [to appear]. <http://www.cs.utep.edu/vladik/2006/tr06-34c.pdf>
- Bertrand, P., and F. Groupil. 2000. Descriptive statistics for symbolic data. Pages 107-124 in *Analysis of Symbolic Data*, H.-H. Bock and E. Diday (eds.), Springer, Berlin.
- Bevington, P.R., and D.K. Robinson. 1992. *Data Reduction and Error Analysis for the Physical Sciences*. WCB McGraw-Hill, Boston.
- Billard, L., and E. Diday. 2000. Regression analysis for interval-valued data. Pages 369-374 in *Data Analysis, Classification and Related Methods*, H.A.L. Kiers, J.-P.

- Rassoon, P.J.F. Groenen and M. Schader (eds.), Springer, Berlin.  
<http://www.stat.uga.edu/faculty/LYNNE/ifcs2.ps>
- Billard, L., and E. Diday. [n.d.]. Symbolic data analysis: definitions and examples. [manuscript]. [http://www.stat.uga.edu/faculty/LYNNE/tr\\_symbolic.pdf](http://www.stat.uga.edu/faculty/LYNNE/tr_symbolic.pdf)
- Bock, H.-H., and E. Diday. 2000. *Analysis of Symbolic Data*. Springer, Berlin.
- Carroll, R.J., D. Ruppert and L.A. Stefanski. 1995. *Measurement Error in Nonlinear Models*. Chapman & Hall, London.
- Chavent, M., F. de A.T. de Carvalho, Y. Lechevallier and R. Verde. 2005. New clustering methods of interval data. [manuscript].
- Chen, L. 1995. Testing the mean of skewed distributions. *Journal of the American Statistical Association* 90: 767-772.
- Cherubini, C., and P. Masi. 2002. Probabilistic and fuzzy reliability analysis in stability assessment. Pages 207- 217 in *Instability Planning and Management: Seeking Sustainable Solutions to Ground Movement Problems*, R.G. McInnes and J. Jakeways (eds.), Thomas Telford, London.
- Chuang, P.H. 2000. Use of fuzzy sets for evaluating shear strength of soils. *Computers and Geotechnics* 17: 425-446.
- Coleman, H.W., and W.G. Steele, Jr. 1999. *Experimentation and Uncertainty Analysis for Engineers*. John Wiley & Sons, New York.
- Corliss, G.F. 1988. How can you apply interval techniques in an industrial setting? Technical Report No. 301. Department of Mathematics, Statistics and Computer Science, Marquette University, Milwaukee, Wisconsin.
- Cormen, Th.H., C.E. Leiserson, R.L. Rivest, and C. Stein. 2001. *Introduction to Algorithms*, MIT Press, Cambridge, Massachusetts.
- Crow, E.L., F.A. Davis, and M.W. Maxfield. 1960. *Statistics Manual With Examples Taken from Ordnance Development*. Dover Publications, New York.
- Cutler, A.N. 2001a. "A history of the speed of light". <http://www.sigma-engineering.co.uk/light/lightindex.shtml>
- Cutler, A.N. 2001b. "Data from Michelson, Pease and Pearson" (1935). <http://www.sigma-engineering.co.uk/light/series.htm>
- Dantsin, E., V. Kreinovich, A. Wolpert and G. Xiang 2006. Population variance under interval uncertainty: a new algorithm. *Reliable Computing* 12(4): 273-280.
- Denker, J.S. 2003. *Physics Documents*, <http://www.av8n.com/physics/> (especially webpage 13, "A discussion of how to report measurement uncertainties").
- Dieck, R.H. 1997. *Measurement Uncertainty Methods and Applications* (second edition). Instrument Society of America, Research Triangle Park, North Carolina.
- Dwyer, P. 1951. *Linear Computations*. John Wiley, New York.
- EPA [U.S. Environmental Protection Agency]. 2000. *Guidance for Data Quality Assessment: Practical Methods for Data Analysis*, EPA QA/G-9, QA00 Update, Office of Environmental Information, Washington, DC.  
<http://www.epa.gov/r10earth/offices/oea/epaqag9b.pdf>
- EPA [U.S. Environmental Protection Agency] 2002. *Calculating Upper Confidence Limits for Exposure Point Concentrations at Hazardous Waste Sites*. OSWER 9285.6-10. Office of Emergency and Remedial Response, Washington, DC.  
<http://www.epa.gov/oswer/riskassessment/pdf/ucl.pdf>. See especially Appendix A.

- Estler, W.T. 1997. A distribution-independent bound on the level of confidence in the result of a measurement. *Journal of Research of the National Institute of Standards and Technology* 102: 587-588. <http://nvl.nist.gov/pub/nistpubs/jres/102/5/j25est.pdf>
- Feigelson, E.D., and P.I. Nelson. 1985. Statistical methods for astronomical data with upper limits. I. Univariate distributions. *The Astrophysical Journal* 293: 192-206.
- Feller, W. 1948. On the Kolmogorov-Smirnov limit theorems for empirical distributions. *Annals of Mathematical Statistics* 19: 177-189.
- Fernández, C., C.J. León, M.F.J. Steel and F.J Vázquez-Polo. 2001. Bayesian analysis of interval data contingent valuation models and pricing policies. Institute of Mathematics and Statistics Technical Report UKC/IMS/01/29. <http://www.kent.ac.uk/ims/personal/mfs/Flsv.pdf>, <http://www.kent.ac.uk/ims/personal/mfs/fig1-3.ps>, <http://www.kent.ac.uk/ims/personal/mfs/fig4-6.ps>, <http://www.kent.ac.uk/ims/personal/mfs/fig7-10.eps>, and <http://www.kent.ac.uk/ims/personal/mfs/fig11-12.eps>
- Fernández, C., C.J. León, M.F.J. Steel and F.J Vázquez-Polo. 2004. Bayesian analysis of interval data contingent valuation models and pricing policies. *Journal of Business and Economic Studies* 22: 431-442. [http://www2.warwick.ac.uk/fac/sci/statistics/staff/academic/steel/steel\\_homepage/techrep/flsvrevfin.ps](http://www2.warwick.ac.uk/fac/sci/statistics/staff/academic/steel/steel_homepage/techrep/flsvrevfin.ps)
- Ferson, S. 1996. Automated quality assurance checks on model structure in ecological risk assessments. *Human and Environmental Risk Assessment* 2:558-569.
- Ferson, S. 1997. Probability bounds analysis. Pages 669-678 in *Computing in Environmental Resource Management*, A. Gertler (ed.), Air and Waste Management Association and the U.S. Environmental Protection Agency, Pittsburgh.
- Ferson, S. 2001. Probability bounds analysis solves the problem of incomplete specification in probabilistic risk and safety assessments. Pages 173-188 in *Risk-Based Decisionmaking in Water Resources IX*, Y.Y. Haimes, D.A. Moser and E.Z. Stakhiv (eds.), American Society of Civil Engineers, Reston, Virginia.
- Ferson, S. 2002. *RAMAS Risk Calc 4.0 Software: Risk Assessment with Uncertain Numbers*. Lewis Publishers, Boca Raton, Florida.
- Ferson, S., and S. Donald. 1998. Probability bounds analysis. Pages 1203-1208 in *Probabilistic Safety Assessment and Management*, A. Mosleh and R.A. Bari (eds.), Springer-Verlag, New York.
- Ferson, S., and L.R. Ginzburg. 1996. Different methods are needed to propagate ignorance and variability. *Reliability Engineering and Systems Safety* 54: 133-144.
- Ferson, S., and J.G. Hajagos. 2004. Arithmetic with uncertain numbers: rigorous and (often) best-possible answers. *Reliability Engineering and System Safety* 85: 135-152.
- Ferson, S., and T.F. Long. 1995. Conservative uncertainty propagation in environmental risk assessments. Pages 97-110 in *Environmental Toxicology and Risk Assessment*, Third Volume, ASTM STP 1218, J.S. Hughes, G.R. Biddinger and E. Mones (eds.), American Society of Testing and Materials, Philadelphia.
- Ferson, S., and T.F. Long. 1998. Deconvolution can reduce uncertainty in risk analyses. *Risk Assessment: Measurement and Logic*, M. Newman and C. Strojan (eds.), Ann Arbor Press, Ann Arbor, Michigan.
- Ferson, S., L. Ginzburg, V. Kreinovich, L. Longpré and M. Aviles. 2002a. Computing variance for interval data is NP-hard. *ACM SIGACT News* 33 (2): 108-118. <http://www.cs.utep.edu/vladik/2002/tr02-13b.pdf>

- Ferson, S., L. Ginzburg, V. Kreinovich and M. Aviles. 2002b. Exact bounds on sample variance of interval data. Pages 67-69 in *Extended Abstracts of the 2002 SIAM Workshop on Validated Computing*, Toronto. <http://www.cs.utep.edu/vladik/2002/tr02-13.pdf>
- Ferson, S., V. Kreinovich, L. Ginzburg, K. Sentz and D.S. Myers. 2003. *Constructing probability boxes and Dempster-Shafer structures*. Sandia National Laboratories, Technical Report SAND2002-4015, Albuquerque, New Mexico, 2002. Available at <http://www.sandia.gov/epistemic/Reports/SAND2002-4015.pdf> and <http://www.ramas.com/unabridged.zip>
- Ferson, S., R.B. Nelsen, J. Hajagos, D.J. Berleant, J. Zhang, W.T. Tucker, L.R. Ginzburg and W.L. Oberkampf. 2004a. *Dependence in Probabilistic Modeling, Dempster-Shafer Theory, and Probability Bounds Analysis*. SAND2004-3072, Sandia National Laboratories, Albuquerque, New Mexico. <http://www.ramas.com/depend.zip>
- Ferson, S., C.A. Joslyn, J.C. Helton, W.L. Oberkampf and K. Sentz. 2004b. Summary from the epistemic uncertainty workshop: consensus amid diversity. *Reliability Engineering and System Safety* 85: 355-370.
- Ferson, S., L. Ginzburg, V. Kreinovich, L. Longpré and M. Aviles. 2005a. Exact bounds on finite populations of interval data. *Reliable Computing* 11 (3): 207-233. <http://www.cs.utep.edu/vladik/2002/tr02-13f.pdf>
- Ferson, S., D. Myers and D. Berleant. 2005b. *Distribution-free Risk Analysis: I. Range, Mean, and Variance*. Applied Biomathematics Technical Report.
- Fetz, T., M. Oberguggenberger and S. Pittschman. 2000. Applications of possibility and evidence theory in civil engineering. *International Journal of Uncertainty, Fuzziness and Knowledge-based Systems* 8: 295-309.
- Fetz, T., M. Oberguggenberger, J. Jäger, D. Köll, G. Krenn and H. Lessmann. 1999. Fuzzy models in geotechnical engineering and construction management. *Computer-aided Civil and Infrastructure Engineering* 14: 93-106.
- Frank, M.J., R.B. Nelsen and B. Schweizer. 1987. Best-possible bounds for the distribution of a sum—a problem of Kolmogorov. *Probability Theory and Related Fields* 74: 199-211.
- Fréchet, M., 1935. Généralisations du théorème des probabilités totales. *Fundamenta Mathematica* 25: 379–387.
- Fréchet, M. 1951. Sur les tableaux de corrélation dont les marges sont données. *Ann. Univ. Lyon, Sect. A* 9: 53-77.
- Fylstra, D., L. Lason, J. Watson and A. Waren. 1998. Design and use of the Microsoft Excel Solver. *Interfaces* 28 (5 September-October): 29-55.
- Garey, M.E., and D. S. Johnson. 1979. *Computers and Intractability: A Guide to the Theory of NP-completeness*. Freeman, San Francisco.
- Gilbert, R.O. 1987. *Statistical Methods for Environmental Pollution Monitoring*. Van Nostrand Reinhold, New York.
- Gioia, F., and C.N. Lauro. 2005. Basic statistical methods for interval data. *Statistica Applicata [Italian Journal of Applied Statistics]* 17(1): 75-104.
- Gioia, F., and C.N. Lauro. 2006. Principal component analysis for interval data. *Computational Statistics* 21(2) 343-363.
- Goos, G., and J. Hartmanis. 1975. *Interval Mathematics*. Lecture Notes in Computer Science, Springer-Verlag, Berlin.

- Granvilliers, L., V. Kreinovich, and N. Müller. 2004. Novel approaches to numerical software with result verification. Pages 274-305 in *Numerical Software with Result Verification*, R. Alt, A. Frommer, R.B. Kearfott, and W. Luther (eds.), International Dagstuhl Seminar, Dagstuhl Castle, Germany, Springer Lectures Notes in Computer Science, volume 2991, Springer, New York.
- Graybill, F.A., and R.B. Deal. 1959. Combining unbiased estimators. *Biometrics* 15: 543-550.
- Grosof, B.N. 1986. An inequality paradigm for probabilistic knowledge: the logic of conditional probability intervals. *Uncertainty in Artificial Intelligence*. L.N. Kanal and J.F. Lemmer (eds.), Elsevier Science Publishers, Amsterdam.
- Hailperin, T. 1986 *Boole's Logic and Probability*. North-Holland, Amsterdam.
- Hajagos, J. 2005. *Accurately Computing Ecological Risk under Measurement Uncertainty*. Dissertation, Stony Brook University, Stony Brook, New York.
- Hall, P. 1988. Theoretical comparison of bootstrap confidence intervals. *Annals of Statistics* 16: 927-953.
- Hall, P. 1992. On the removal of skewness by transformation. *Journal of the Royal Statistical Society B* 54: 221-228.
- Helsel, D.R. 1990. Less than obvious: Statistical treatment of data below the detection limit. *Environmental Science and Technology* 24: 1766-1774.
- Helsel, D.R. 2005. *Nondetects and Data Analysis: Statistics for Censored Environmental Data*. Wiley, New York.
- Henrion, M., and B. Fischhoff. 1986. Assessing uncertainty in physical constants. *American Journal of Physics* 54(9): 791-798.
- Hoffman, F.O., and J.S. Hammonds. 1994. Propagation of uncertainty in risk assessments: The need to distinguish between uncertainty due to lack of knowledge and uncertainty due to variability. *Risk Analysis* 14: 707-712.
- Horowitz, J., and C. Manski. 2000. Nonparametric analysis of randomized experiments with missing covariate and outcome data. *Journal of the American Statistical Association* 95: 77-84.
- Huang, Y.-T., and T.J. Siller. 1997. Fuzzy representation and reasoning in geotechnical site characterization. *Computers and Geotechnics* 21: 65-86.
- Huber, P.J. 1981. *Robust Statistics*. Wiley, New York. Republished in 2004.
- Huber, W. [Quantitative Decisions]. 2001. "Statistical tests: the Chebyshev UCL proposal". [http://www.quantdec.com/envstats/notes/class\\_12/ucl.htm#Chebyshev](http://www.quantdec.com/envstats/notes/class_12/ucl.htm#Chebyshev)
- Insua, D.R., and F. Ruggeri. 2000. *Robust Bayesian Analysis*. Lecture Notes in Statistics, Volume 152. Springer-Verlag, New York.
- ISO [International Organization for Standardization]. 1993. *Guide to the Expression of Uncertainty in Measurement*. International Organization for Standardization, Geneva, Switzerland.
- Isobe, T., E.D. Feigelson and P.I. Nelson. 1986. Statistical methods for astronomical data with upper limits. II. Correlation and regression. *The Astrophysical Journal* 306: 490-507.
- Jájá, J. 1992. *An Introduction to Parallel Algorithms*. Addison-Wesley, Reading, Massachusetts.
- JCGM [Joint Committee for Guides in Metrology]. 2006. Evaluation of measurement data — Supplement 1 to the "Guide to the expression of uncertainty in



- measurement” — Propagation of distributions using a Monte Carlo method.  
Available at [http://www.intermet.jp/JCGM/0610news/Supplement\\_to\\_GUM.pdf](http://www.intermet.jp/JCGM/0610news/Supplement_to_GUM.pdf).
- Jaulin, L., M. Kieffer, O. Didrit, and E. Walter. 2001. *Applied Interval Analysis*. Springer-Verlag, London.
- Johnson, N.J. 1978. Modified  $t$ -tests and confidence intervals for asymmetrical populations. *The American Statistician* 73:536-544.
- Joslyn, C. 1997. Measurement of possibilistic histograms from interval data. *International Journal of General Systems* 26 (1-2): 9-33.
- Kaars, R. 1981. Rapport du groupe de travail sur l’expression des incertitudes au Comité International des Poids et Mesures. *Proc.-Verb. Com. Int. Poids et Mesures* 49: A1-A12. See also Giacomo, P. 1981. News from the BIPM. *Metrologia* 17: 69-74. See also <http://physics.nist.gov/cuu/Uncertainty/international1.html>.
- Killeen, P.R. 2005. Replicability, confidence, and priors. *Psychological Science* 16(12): 1009-1012.
- Klir, G., and B. Yuan. 1995. *Fuzzy Sets and Fuzzy Logic: Theory and Applications*. Prentice Hall, Upper Saddle River, NJ.
- Kolmogorov, [Kolmogoroff] A. 1941. Confidence limits for an unknown distribution function. *Annals of Mathematical Statistics* 12: 461-463.
- Kreinovich, V. 2004. Probabilities, intervals, what next? Optimization problems related to extension of interval computations to situations with partial information about probabilities. *Journal of Global Optimization* 29 (3): 265-280.
- Kreinovich, V., and L. Longpré. 2003. Computational complexity and feasibility of data processing and interval computations, with extension to cases when we have partial information about probabilities. Pages 19-54 in *Proceedings of the Conference on Computability and Complexity in Analysis CCA2003*, V. Brattka, M. Schroeder, K. Weihrauch and N. Zhong (eds.).
- Kreinovich, V., and L. Longpré. 2004. Fast quantum algorithms for handling probabilistic and interval uncertainty. *Mathematical Logic Quarterly* 50 (4/5): 507-518.
- Kreinovich, V., A. Lakeyev, J. Rohn, and P. Kahl. 1997. *Computation Complexity and Feasibility of Data processing and Interval Computations*. Kluwer, Dordrecht.
- Kreinovich, V., P. Patangay, L. Longpré, S. A. Starks, C. Campos, S. Ferson, and L. Ginzburg. 2003. Outlier detection under interval and fuzzy uncertainty: algorithmic solvability and computational complexity. Pages 401-406 in *Proceedings of the 22nd International Conference of the North American Fuzzy Information Processing Society NAFIPS 2003*.
- Kreinovich, V., L. Longpré, S. Ferson, and L. Ginzburg. 2004a. *Computing Higher Central Moments for Interval Data*. University of Texas at El Paso, Department of Computer Science, Technical Report UTEP-CS-03-14b. <http://www.cs.utep.edu/vladik/2003/tr03-14b.pdf>
- Kreinovich, V., L. Longpré, P. Patangay, S. Ferson, and L. Ginzburg. 2004b. Outlier detection under interval uncertainty: algorithmic solvability and computational complexity. Pages 238-245 in *Large-Scale Scientific Computing*, I. Lirkov, S. Margenov, J. Wasniewski and P. Yalamov (eds.), Proceedings of the 4th International Conference LSSC 2003, Springer Lecture Notes in Computer Science, volume 2907, Springer, New York.

- Kreinovich, V., L. Longpré, P. Patangay, S. Ferson and L. Ginzburg. 2005a. Outlier detection under interval uncertainty: algorithmic solvability and computational complexity. *Reliable Computing* 11 (1): 59-76. <http://www.cs.utep.edu/vladik/2003/tr03-10f.pdf>
- Kreinovich, V., H.T. Nguyen and B. Wu. [2005b]. On-line algorithms for computing mean and variance of interval data, and their use in intelligent systems. *Information Sciences* [to appear]. <http://www.cs.utep.edu/vladik/2003/tr03-24d.pdf>
- Kreinovich, V., G. Xiang, S. A. Starks, L. Longpré, M. Ceberio, R. Araiza, J. Beck, R. Kandathi, A. Nayak, R. Torres and J. Hajagos. 2006a. Towards combining probabilistic and interval uncertainty in engineering calculations: algorithms for computing statistics under interval uncertainty, and their computational complexity. *Reliable Computing* 12 (6): 471-501. <http://www.cs.utep.edu/vladik/2004/tr04-20b.pdf>
- Kreinovich, V., G. Xiang and S. Ferson. 2006b. Computing mean and variance under Dempster-Shafer uncertainty: towards faster algorithms. *International Journal of Approximate Reasoning* 42: 212-227. <http://www.cd.utep.edu/vladik/2005/tr05-24a.pdf>
- Kutterer, H. 2004. Statistical hypothesis tests in case of imprecise data. Pages 49-56 in *Proceedings of the Fifth Hotine-Marussi Symposium*, International Association of Geodesy Symposia, Springer, Berlin.
- Kuznetsov, V. P. 1991. *Interval Statistical Models* (in Russian). Radio i Svyaz, Moscow.
- Laplace, Marquis de, P.S. 1820. *Théorie analytique de probabilités* (edition troisième). Courcier, Paris. The introduction (*Essai philosophique sur les probabilités*) is available in an English translation in *A Philosophical Essay on Probabilities* (1951), Dover Publications, New York.
- Land, C.E. 1972. An evaluation of approximate confidence interval estimation methods for lognormal means. *Technometrics* 14: 145-158.
- Land, C. E. 1975. Tables of confidence limits for linear functions of the normal mean and variance. *Selected Tables in Mathematical Statistics*, Vol III, p 385-419.
- Lauro, N.C., and F. Palumbo. 2000. Principal components analysis of interval data: a symbolic data analysis approach. *Computational Statistics* 15 (1): 73-87. [http://www.dms.unina.it/dati/pdf/lauro/Lauro\\_Palumbo.pdf](http://www.dms.unina.it/dati/pdf/lauro/Lauro_Palumbo.pdf)
- Lauro, C.N., and F. Palumbo. 2003. Some results and new perspectives in Principal Component Analysis for interval data. Pages 237-244 Atti del Convegno CLADAG'03 Gruppo di Classificazione della Società Italiana di Statistica. <http://www.economiamc.org/repo/39/CLADAG.pdf>
- Lauro, C.N., and F. Palumbo. 2005. Principal component analysis for non-precise data. Pages 173–184 in *New Developments in Classification and Data Analysis*, M. Vichi et al. (eds.), Springer.
- Lauro, C.N., R. Verde and A. Irpino. 2007. Principal component analysis of symbolic data described by intervals. *Symbolic Data Analysis and the SODAS Software*, E. Diday and M. Noirhome Fraiture, John Wiley and Sons.
- Levenson, M.S., D.L. Banks, K.R. Eberhardt, L.M. Gill, W.F. Guthrie, H.K. Liu, M.G. Vangel, J.H. Yen, and N.F. Zhang. 2000. An approach to combining results from multiple methods motivated by the ISO GUM. *Journal of Research of the National Institute of Standards and Technology* 105 (4): 571-579. Available at <http://nvl.nist.gov/pub/nistpubs/jres/105/4/j54lev.pdf>.

- Little, R.J.A., and D.B. Rubin. 1987. *Statistical Analysis with Missing Data*. Wiley, New York.
- Lodwick, W.A., and K.D. Jamison. 2003. Estimating and validating the cumulative distribution of a function of random variables: toward the development of distribution arithmetic. *Reliable Computing* 9 (2): 127-141.
- Manes, E.G. 1982. A class of fuzzy theories. *Journal of Mathematical Analysis and Its Applications* 85: 409-451.
- Manski, C.F. 2003. *Partial Identification of Probability Distributions*, Springer Series in Statistics, Springer, New York.
- Manski, C., and E. Tamer 2002. Inference on regressions with interval data on a regressor or outcome. *Econometrica* 70 (2): 519-546.
- Manton, K.G., M.A. Woodbury and H.D. Tolley. 1994. *Statistical Applications using Fuzzy Sets*. Wiley, New York.
- Marino, M., and F. Palumbo. 2002. Interval arithmetic for the evaluation of imprecise data effects in least squares linear regression. *Statistica Applicata* 14 (3): 277-291.
- Marino, M., and F. Palumbo. 2003. Interval linear regression: an application to soil permeability analysis. *Analisi statistica multivariata per le Scienze Economico-Sociali, le Scienze Naturali e la Tecnologia*, Convegno Intermedio della Società Italiana di statistica. [http://www.economiamc.org/repo/39/Marino\\_Palumbo\\_rev.pdf](http://www.economiamc.org/repo/39/Marino_Palumbo_rev.pdf)
- Markov, S. 1990. Least-square approximations under interval input data. Pages 133-147 in *Contributions to Computer Arithmetic and Self-validating Numerical Methods*, C. Ullrich (ed.), J.C. Baltzer AG, Scientific Publishing.
- Martinez, M., L. Longpré, V. Kreinovich, S. A. Starks, and H. T. Nguyen. 2003. Fast quantum algorithms for handling probabilistic, interval, and fuzzy uncertainty. Pages 395-400 in *Proceedings of the 22nd International Conference of the North American Fuzzy Information Processing Society NAFIPS 2003*, Chicago, Illinois.
- Meeker, W.Q., and L.A. Escobar. 1995. *Statistical Methods for Reliability Data*. John Wiley and Sons, New York.
- Meier, P. 1953. Variance of a weighted mean. *Biometrics* 9: 59-73.
- Michelson, A.A., F.G. Pease and F. Pearson. 1935. Measurement of the velocity of light in a partial vacuum. *Astronomical Journal* 82: 26. See Cutler (2001b).
- Miller, L.H. 1956. Table of percentage points of Kolmogorov statistics. *Journal of the American Statistical Association* 51: 111-121.
- Möller, B., and M. Beer. 2004. *Fuzzy Randomness — Uncertainty in Civil Engineering and Computational Mechanics: A Comprehensive Consideration of Uncertainty in the Numerical Analysis, the Safety Assessment, and the Design of Structures*. Springer.
- Möller, B., and W. Reuter. 2006. Prediction of uncertain structural responses with fuzzy time series. Pages 419-437 in *Proceedings of the NSF Workshop on Reliable Engineering Computing: Modeling Errors and Uncertainty in Engineering Computations*, R.L. Muhanna and R.L. Mullen (eds.), Savannah, Georgia. [http://www.gtrep.gatech.edu/workshop/rec06/papers/Moeller\\_paper.pdf](http://www.gtrep.gatech.edu/workshop/rec06/papers/Moeller_paper.pdf)
- Mood, A.M., F.A. Graybill and D.C. Boes. 1974. *Introduction to the Theory of Statistics*. McGraw-Hill, New York.
- Moore, R.E. 1966. *Interval Analysis*. Prentice Hall, Englewood Cliffs, New Jersey.
- Moore, R.E. 1979. *Methods and Applications of Interval Analysis*. SIAM, Philadelphia.

- Moore, R.E., and W.A. Lodwick. 2003. Interval Analysis and Fuzzy Set Theory. *Fuzzy Sets and Systems* 135 (1): 5-9.
- Morgan, M.G., and M. Henrion. 1990. *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*. Cambridge University Press, Cambridge.
- Morgenstein, D., and V. Kreinovich. 1995. Which algorithms are feasible and which are not depends on the geometry of space-time. *Geombinatorics* 4 (3): 80-97.
- Neumaier, A. 1990. *Interval Methods for Systems of Equations*. Cambridge University Press, Cambridge.
- Neumann, I., H. Kutterer and S. Schön. 2006. Outlier detection in geodetic applications with respect to observation imprecision. Pages 75-90 in *Proceedings of the NSF Workshop on Reliable Engineering Computing: Modeling Errors and Uncertainty in Engineering Computations*, R.L. Muhanna and R.L. Mullen (eds.), Savannah, Georgia, [http://www.gtrep.gatech.edu/workshop/rec06/papers/Neumann\\_paper.pdf](http://www.gtrep.gatech.edu/workshop/rec06/papers/Neumann_paper.pdf).
- Nguyen, H.T., and V. Kreinovich. 1996. Nested Intervals and Sets: Concepts, Relations to Fuzzy Sets, and Applications. Pages 245-290 in *Applications of Interval Computations*, R.B. Kearfott and V. Kreinovich, Kluwer, Dordrecht.
- Nguyen, H.T., and E. A. Walker. 1999. *First Course in Fuzzy Logic*. CRC Press, Boca Raton, Florida, 1999.
- Nguyen, H.T., B. Wu, and V. Kreinovich. 2000. Shadows of fuzzy sets—a natural approach towards describing 2-D and multi-D fuzzy uncertainty in linguistic terms. Pages 340-345 in *Proceedings of the 9th IEEE International Conference on Fuzzy Systems FUZZ-IEEE 2000*, Vol. 1, San Antonio, Texas.
- Nguyen, H.T., T. Wang, and V. Kreinovich. 2003. Towards foundations of processing imprecise data: from traditional statistical techniques of processing crisp data to statistical processing of fuzzy data. Pages 895-900 in *Proceedings of the International Conference on Fuzzy Information Processing: Theories and Applications FIP'2003*, Vol. II, Y. Liu, G. Chen, M. Ying, and K.-Y. Cai (eds.), Beijing, China.
- Nivlet, P., F. Fournier, and J. Royer. 2001a. A new methodology to account for uncertainties in 4-D seismic interpretation. Pages 1644-1647 in *Proceedings of the 71st Annual International Meeting of the Society of Exploratory Geophysics SEG'2001*, San Antonio, Texas.
- Nivlet, P., F. Fournier, and J. Royer. 2001b. Propagating interval uncertainties in supervised pattern recognition for reservoir characterization. Paper SPE-71327 in *Proceedings of the 2001 Society of Petroleum Engineers Annual Conference SPE 2001*, New Orleans.
- Oliverio, P. 1996. Self-generating Pythagorean quadruples and  $n$ -tuples. *Fibonacci Quarterly* (May 1996): 98-101.
- Osegueda, R., V. Kreinovich, L. Potluri, R. Aló. 2002. Non-destructive testing of aerospace structures: granularity and data mining approach. Pages 685-689 in *Proceedings of FUZZ-IEEE 2002*, Vol. 1, Honolulu, Hawaii.
- Palumbo, F., A. Iripino. 2005. Multidimensional interval-data: metrics and factorial analysis. *Proceedings ASMDA 2005*, Brest. <http://www.economiamec.org/repo/39/689.pdf> or <http://asmda2005.enst-bretagne.fr/IMG/pdf/proceedings/689.pdf>

- Pericchi, L.R. 2000. Sets of prior probabilities and Bayesian robustness. Imprecise Probability Project. <http://ippserv.rug.ac.be/documentation/robust/robust.html>, <http://www.sipta.org/documentation/robust/pericchi.pdf>
- Pericchi, L.R., and P. Walley. 1991. Robust Bayesian credible intervals and prior ignorance. *International Statistical Review* 58: 1-23.
- Plous, S. 1993. *The Psychology of Judgment and Decision Making*. McGraw-Hill, New York.
- Rabinovich, S. 1993. *Measurement Errors: Theory and Practice*, American Institute of Physics, New York.
- Regan, H., S. Ferson, and D. Berleant. 2004. Equivalence of five methods for bounding uncertainty. *Journal of Approximate Reasoning* 36 (1): 1-30.
- Römer, C., and A. Kandel. 1995. Statistical tests for fuzzy data. *Fuzzy Sets and Systems* 72: 1-26.
- Rousseeuw, P., and A. Leroy. 1987. *Robust Regression and Outlier Detection*. Wiley.
- Rowe, N. 1988. Absolute bounds on the mean and standard deviation of transformed data for constant-sign-derivative transformations. *SIAM Journal of Scientific and Statistical Computing* 9:1098–1113.
- Rubio, E., J.W. Hall and M.G. Anderson 2004. Uncertainty analysis in a slope hydrology and stability model using probabilistic and imprecise information. *Computers and Geotechnics* 31: 529-536.
- Saila, S.B., and S. Ferson. 1998. Fuzzy regression in fisheries science—some methods and applications. *Fishery Stock Assessment Models*. F. Funk, T.J. Quinn II, J. Heifetz, J.N. Ianelli, J.E. Powers, J.F. Schweigert, P.J. Sullivan, and C.-I. Zhang (eds.), AK-SG-98-01, Alaska Sea Grant College Program, Fairbanks, Alaska.
- Saltelli, A., Chan, K., Scott, E.M., (eds.). 2001. *Sensitivity Analysis*. John Wiley, New York.
- Saw, J.G., M.C.K. Yang and T.C. Mo. 1984. Chebyshev inequality with estimated mean and variance. *The American Statistician* 38: 130-132.
- Saw, J.G., M.C.K. Yang and T.C. Mo. 1984. Corrections. *The American Statistician* 42: 166.
- Schmitt, J.H.M.M. 1985. Statistical analysis of astronomical data containing upper bounds: general methods and examples drawn from X-ray astronomy. *The Astrophysical Journal* 293: 178-191.
- Schulz, T.W., and S. Griffin 1999. Estimating risk assessment exposure point concentrations when the data are not normal or lognormal. *Risk Analysis* 19:577-584.
- Sen, P.K. 1968. Estimates of regression coefficient based on Kendall's tau. *Journal of the American Statistical Association* 63(324): 1379-1389.
- Shewart, W.A. 1939. *Statistical Method from the Viewpoint of Quality Control*. Graduate School, Department of Agriculture, Washington, DC.
- Shafer, G. 1976. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, New Jersey.
- Shlyakhter, A.I. 1993. Statistics of past errors as a source of safety factors for current models. *Model Uncertainty: Its Characterization and Quantification*, A. Mosleh, N. Siu, C. Smidts, and C. Lui (eds.). Center for Reliability Engineering, University of Maryland College Park, Maryland.

- Shlyakhter, A.I. 1994. Uncertainty estimates in scientific models: lessons from trends in physical measurements, population and energy projections. *Uncertainty Modeling and Analysis: Theory and Applications*, B.M. Ayyub and M.M. Gupta (eds.), North-Holland-Elsevier Scientific Publishers. Chapter available at [http://sdg.lcs.mit.edu/~ilya\\_shl/alex/94c\\_uncertainty\\_scientific\\_models\\_physical\\_measurements\\_projections.pdf](http://sdg.lcs.mit.edu/~ilya_shl/alex/94c_uncertainty_scientific_models_physical_measurements_projections.pdf).
- Shmulevich, I., and W. Zhang. 2002. Binary analysis and optimization-based normalization of gene expression data. *Bioinformatics* 18 (4): 555-565.
- Siegel, A.F. 1982. Robust regression using repeated medians. *Biometrika* 69: 242-244.
- Singh, A.K., A. Singh and M. Engelhardt. 1997. The lognormal distribution in environmental applications. EPA/600/R-97/006.
- Singh, A.K., A. Singh, M. Engelhardt and J. M. Nocerino. 2000. On the computation of the upper confidence limit of the mean of contaminant data distributions. Prepared for the Office of Research and Development, Technical Support Center, U.S. Environmental Protection Agency, Las Vegas, Nevada.
- Smirnov [Smirnoff], N. 1939. On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Bulletin de l'Université de Moscou, Série internationale (Mathématiques)* 2: (fasc. 2).
- Sokal, R.R., and F.J. Rohlf. 1981. *Biometry*. Freeman, New York.
- Springer, M.D. 1979. *The Algebra of Random Variables*. Wiley, New York.
- Starks, S.A., V. Kreinovich, L. Longpré, M. Ceberio, G. Xiang, R. Araiza, J. Beck, R. Kandathi, A. Nayak and R. Torres. 2004. Towards integration of probabilistic and interval errors in engineering calculations. Pages 193-213 in *Proceedings of the Workshop on Reliable Engineering Computing*, Savannah, Georgia. <http://www.cs.utep.edu/vladik/2004/tr04-20.pdf> [related slide presentation [http://www.gtrep.gatech.edu/rec/recworkshop/presentations/Kreinovich\\_2.pdf](http://www.gtrep.gatech.edu/rec/recworkshop/presentations/Kreinovich_2.pdf)]
- Stigler, S.M. 1996. Statistics and the question of standards. *Journal of Research of the National Institute of Standards and Technology* 101: 779-789.
- Stuart, A., and J.K. Ord 1994. *Kendall's Advanced Theory of Statistics*. Edward Arnold, London.
- Student [W.S. Gosset] 1908. On the probable error of the mean. *Biometrika* 6: 1-25.
- Taylor, B.N., and C.E. Kuyatt. 1994. *Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results*. NIST Technical Note 1297, National Institute of Standards and Technology, Washington, DC. Available on-line at <http://physics.nist.gov/Pubs/guidelines/contents.html>. See also web guidance at <http://physics.nist.gov/cuu/Uncertainty/index.html>.
- Theil, H. 1950. A rank-invariant method for linear and polynomial regression analysis. *Koninklijke Nederlandse Akademie van Wetenschappen Series A* 53: 386-392 (Part I), 521-525 (Part II), 1397-1412 (Part III).
- Trautman, A. 1998. Pythagorean spinors and penrose twistors. *The Geometric Universe: Science, Geometry, and the Work of Roger Penrose*, S.A. Hugget et al. (eds.), Oxford University Press, Oxford.
- Viertl, R. 1996. *Statistical Methods for Non-precise Data*. CRC Press, Boca Raton.
- Wadsworth, H.M., Jr. (ed.). 1990. *Handbook of Statistical Methods for Engineers and Scientists*. McGraw-Hill Publishing, New York.

- Walley, P. 1991. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London.
- Wang, X.Q., and Q. Yu. 2005. Unbiasedness of the Theil-Sen estimator. *Journal of Non-parametric Statistics* 685-695. Related slide presentation by Qiqing Yu available at <http://www.stat.sc.edu/~nglenn/shuangLiTalk.ppt>.
- Ward, M.P., and M.C. Jones. 1995. *Kernel Smoothing*. Chapman and Hall, London.
- Wilks, S.S. 1962. *Mathematical Statistics*. John Wiley & Sons, New York.
- Williamson, R.C. 1989. *Probabilistic Arithmetic*. Ph.D. dissertation, University of Queensland, Australia. <http://theorem.anu.edu.au/~williams/papers/thesis300dpi.ps>
- Williamson, R.C., and T. Downs 1990. Probabilistic arithmetic I: Numerical methods for calculating convolutions and dependency bounds. *International Journal of Approximate Reasoning* 4: 89-158.
- Wise, B.P., and M. Henrion. 1986. A framework for comparing uncertain inference systems to probability. *Uncertainty in Artificial Intelligence*, L.N. Kanal and J.F. Lemmer (eds.), Elsevier Science Publishers, Amsterdam.
- Wong, A. 1993. A note on inference for the mean parameter of the gamma distribution. *Stat. Prob. Lett.* 17: 61-66.
- Wu, B., H.T. Nguyen and V. Kreinovich. 2003. Real-time algorithms for statistical analysis of interval data. Pages 483-490 in *Proceedings of the International Conference on Information Technology InTech'03*, Chiang Mai, Thailand. <http://www.cs.utep.edu/vladik/2003/tr03-24a.pdf>
- Xiang, G. 2006. Fast algorithm for computing the upper endpoint of sample variance for interval data: case of sufficiently accurate measurements. *Reliable Computing* 12 (1): 59-64. <http://www.cs.utep.edu/vladik/2004/tr04-31.pdf>
- Xiang, G., S.A. Starks, V. Kreinovich and L. Longpré. 2006. New algorithms for statistical analysis of interval data. *Springer Lecture Notes in Computer Science* 3732: 189-196. <http://www.cs.utep.edu/vladik/2004/tr04-04b.pdf>
- Xiang, G., M. Ceberio and V. Kreinovich. 2007a. Computing population variance and entropy under interval uncertainty: linear-time algorithms. [to appear]. <http://www.cs.utep.edu/vladik/2006/tr06-28.pdf>
- Xiang, G., V. Kreinovich and S. Ferson. 2007b. Fitting a normal distribution to interval and fuzzy data. *Proceedings of the 26th International Conference of the North American Fuzzy Information Processing Society NAFIPS'2007*, San Diego, California, June 24-27, 2007. <http://www.cs.utep.edu/vladik/2007/tr07-21.pdf>
- Yager, R.R. 1986. Arithmetic and other operations on Dempster-Shafer structures. *International Journal of Man-machine Studies* 25: 357-366.
- Youden, W.J. 1972. Enduring values. *Technometrics* 14:1-11.
- Young, R.C. 1931. The algebra of multi-valued quantities. *Mathematische Annalen* 104: 260-290. <http://www.cs.utep.edu/interval-comp/young.pdf>
- Zaffalon, M. 2002. Exact credal treatment of missing data. *Journal of Statistical Planning and Inference* 105(1): 105-122. See <http://www.idsia.ch/~zaffalon/papers/md.pdf>
- Zaffalon, M. 2005. Conservative rules for predictive inference with incomplete data. ISIPTA'05: Proceedings of the Fourth International Symposium on Imprecise Probabilities and Their Applications. Society for Imprecise Probabilities and Their Applications, Manno, Switzerland. <http://www.idsia.ch/~zaffalon/papers/2005isipta-mistat.pdf>

- Zhang, N.F. 2006. The uncertainty associated with the weighted mean of measurement data. *Metrologia* 43: 195-204.
- Zhang, W., I. Shmulevich, and J. Astola. 2004. *Microarray Quality Control*, Wiley, Hoboken, New Jersey.
- Zhou, X.-H., and Gao, S. 1997. Confidence intervals for the lognormal mean. *Statistics in Medicine* 16: 783-790.
- Zhou, X.-H., and Gao, S. 2000. One-sided confidence intervals for means of positively skewed distributions. *The American Statistician* 54: 100-104.



## External Distribution

Marv Adams  
Department of Nuclear Engineering  
Texas A&M University  
3133 TAMU  
College Station, TX 77843-3133

Bilal Ayyub (2)  
Department of Civil Engineering  
University of Maryland  
College Park, MD 20742-3021

Ivo Babuska  
TICAM  
Mail Code C0200  
University of Texas at Austin  
Austin, TX 78712-1085

S. Balachandar  
Dept. of Mechanical & Aerospace Engr.  
University of Florida  
231 MAE-A, PO Box 116250  
Gainesville, FL 32611-6205

Timothy Barry  
Science-policy, Planning and Evaluation  
USEPA Headquarters  
MC2174  
401 M Street, S.W.  
Washington, DC 20460

John Benek  
AFRL/VAAC  
2210 Eighth St.  
Wright-Patterson AFB, OH 45433

James Berger  
Inst. of Statistics and Decision Science  
Duke University  
Box 90251  
Durham, NC 27708-0251

Vicki M. Bier  
Department of Industrial Engineering  
University of Wisconsin  
Madison, WI 53706

Lynne Billard  
Department of Statistics  
University of Georgia  
Athens, GA 30602-1952

Mark Brandyberry  
Computational Science and Engineering  
2264 Digital Computer Lab, MC-278  
1304 West Springfield Ave.  
University of Illinois  
Urbana, IL 61801

John A. Cafeo  
General Motors R&D Center  
Mail Code 480-106-256  
30500 Mound Road  
Box 9055  
Warren, MI 48090-9055

Wei Chen (2)  
Department of Mechanical Engineering  
Northwestern University  
2145 Sheridan Road, Tech B224  
Evanston, IL 60208-3111

Hugh Coleman  
Department of Mechanical &  
Aero. Engineering  
University of Alabama/Huntsville  
Huntsville, AL 35899

Thomas A. Cruse  
AFRL Chief Technologist  
1981 Monahan Way  
Bldg. 12, Room 107  
Wright-Patterson AFB, OH 45433-7132

Department of Energy (5)  
Attn: Christopher Deeney, NA-113  
Dimitri Kusnezov, NA-114  
Njema Frazier, NA-114  
Kevin Greenaugh, NA-115  
Jamlieh Soudah, NA-116  
Forrestal Building  
1000 Independence Ave., SW  
Washington, DC 20585

Bridgette R. DeShields  
BBL Sciences  
1670 Corporate Circle, Suite 200  
Petaluma, CA 94954-9612

Prof. Urmila Diwekar (2)  
University of Illinois at Chicago  
Chemical Engineering Dept.  
810 S. Clinton St.  
209 CHB, M/C 110  
Chicago, IL 60607

David Dolling  
Aerospace Engineering  
& Engineering Mechanics  
University of Texas at Austin  
Austin, TX 78712-1085

Scott J. Duncan  
Systems Realization Laboratory  
Woodruff School of Mechanical Eng.  
Georgia Institute of Technology  
Atlanta, Georgia 30332-0405

Isaac Elishakoff  
Dept. of Mechanical Engineering  
Florida Atlantic University  
777 Glades Road  
Boca Raton, FL 33431-0991

Ashley Emery  
Dept. of Mechanical Engineering  
Box 352600  
University of Washington  
Seattle, WA 98195-2600

Scott Ferson (20)  
Applied Biomathematics  
100 North Country Road  
Setauket, NY 11733-1345

James J. Filliben (2)  
NIST, Stop 8980  
100 Bureau Drive  
Gaithersburg, MD 20899-8980

Jeffrey T. Fong  
NIST, Stop 8910  
100 Bureau Drive  
Gaithersburg, MD 20899-8910

Mike Giltrud  
Defense Threat Reduction Agency  
Attn: DTRA/CXS  
8725 John J. Kingman Rd  
Mail Stop 6201  
Fort Belvoir, VA 22060-6201

Michael Ginevan  
307 Hamilton Avenue  
Silver Spring, MD 20901

James Glimm (2)  
Dept. of Applied Math & Statistics  
P138A  
State University of New York  
Stony Brook, NY 11794-3600

Melody S. Goodman  
Preventive Medicine  
Health Sciences Center, 3-071  
Stony Brook, NY 11794-8338

James Gran  
SRI International  
Poulter Laboratory AH253  
333 Ravenswood Avenue  
Menlo Park, CA 94025

John W. Green  
DuPont Stine-Haskell Research Center  
1090 Elkton Road  
P.O. Box 50  
BLDG H-1 / 423B  
Newark, DE 19714-0050

Yacov Haimes  
Systems and Information Engineering  
University of Virginia  
P.O. Box 400747  
151 Engineer's Way  
Charlottesville, VA 22904

Dale Hattis  
CENTED  
Clark University  
950 Main Street  
Worcester, MA 01610

Richard Hills (2)  
Mechanical Engineering Dept.  
New Mexico State University  
P. O. Box 30001/Dept. 3450  
Las Cruces, NM 88003-8001

F. Owen Hoffman (2)  
SENES  
102 Donner Drive  
Oak Ridge, TN 37830

Luc Huyse  
Reliability & Materials Integrity, B-128  
Southwest Research Institute  
P.O. Drawer 28510  
San Antonio, TX 78228-0510

Gianluca Iaccarino  
Center for Turbulence Research  
Building 500, Room 500A  
488 Escondido Mall,  
Stanford University  
Stanford, CA 94305-3035

Mark Inlow  
Department of Mathematics  
Rose-Hulman Institute of Technology  
5500 Wabash Avenue  
Terre Haute, IN 47803

Mark Jablonowski  
Post Office Box 521  
Tariffville, CT 06081

Alan Karr  
Inst. of Statistics and Decision Science  
Duke University  
Box 90251  
Durham, NC 27708-0251

Vladik Kreinovich (2)  
Department of Computer Science  
University of Texas at El Paso  
500 W. University  
El Paso, TX 79968

Bryce Landenberger  
Health & Environmental Research  
The Dow Company  
1803 Building  
Midland, MI 48674

John Lehner  
Energy Sciences and Technology  
Building 475, Mailstop 475C  
Brookhaven National Laboratory  
Upton, NY 11973-5000

Robert Lust  
General Motors, R&D and Planning  
MC 480-106-256  
30500 Mound Road  
Warren, MI 48090-9055

Sankaran Mahadevan (2)  
Civil & Environmental Engineering  
Vanderbilt University  
Box 6077, Station B  
Nashville, TN 37235

Hans Mair  
Institute for Defense Analysis  
Operational Evaluation Division  
4850 Mark Center Drive  
Alexandria VA 22311-1882

Charles F. Manski  
Department of Economics  
Northwestern University  
2001 Sheridan Road  
Evanston, IL 60208

Keith D. McCroan  
National Air and Radiation  
Environmental Laboratory  
540 South Morris Avenue  
Montgomery, AL 36115-2601

Richard McGrath  
Weston Solutions  
10 Lyman Street  
Pittsfield, MA 01201

Gregory McRae (2)  
Dept. of Chemical Engineering  
Massachusetts Institute of Technology  
Cambridge, MA 02139

John G. Michopoulos  
Naval Research Laboratory,  
Special Projects Group, Code 6303  
Computational Multiphysics Systems Lab  
Washington DC 20375, USA

Granger Morgan  
Carnegie-Mellon University  
Dept. of Engineering & Public Policy  
Pittsburgh, PA 15213

Max Morris (2)  
Department of Statistics  
Iowa State University  
304A Snedecor-Hall  
Ames, IA 50011-1210

Rafi Muhanna  
Regional Engineering Program  
Georgia Tech  
210 Technology Circle  
Savannah, GA 31407-3039

Steven Munch  
Marine Sciences Research Center  
State University of New York  
Stony Brook, NY 11794

NASA/Ames Research Center (2)  
Attn: Unmeel Mehta, MS 229-3  
David Thompson, MS 269-1  
Moffett Field, CA 94035-1000

NASA/Glenn Research Center (2)  
Attn: John Slater, MS 86-7  
Chris Steffen, MS 5-11  
21000 Brookpark Road  
Cleveland, OH 44135

NASA/Langley Research Center (9)  
Attn: Dick DeLoach, MS 236  
Michael Hemsch, MS 499  
Jim Luckring, MS 286  
Joe Morrison, MS 128  
Ahmed Noor, MS 369  
Sharon Padula, MS 159  
Thomas Zang, MS 449  
Lawrence L. Green, MS 462  
Hampton, VA 23681-0001

Robert Nelson  
Dept. of Aerospace & Mechanical Engr.  
University of Notre Dame  
Notre Dame, IN 46556

Roger B. Nelsen  
Dept. of Mathematical Sciences  
Lewis & Clark College  
0615 SW Palatine Hill Road  
Portland OR 97219-7899

Efstratios Nikolaidis (2)  
MIME Dept.  
4035 Nitschke Hall  
University of Toledo  
Toledo, OH 43606-3390

Tinsley Oden (2)  
TICAM  
Mail Code C0200  
University of Texas at Austin  
Austin, TX 78712-1085

Michael Ortiz (2)  
Graduate Aeronautical Laboratories  
California Institute of Technology  
1200 E. California Blvd./MS 105-50  
Pasadena, CA 91125

Chris Paredis  
School of Mechanical Engineering  
Georgia Institute of Technology  
813 Ferst Drive, MARC Rm. 256  
Atlanta, GA 30332-0405

Chris L. Pettit  
Aerospace Engineering Dept.  
MS-11B  
590 Holloway Rd.  
Annapolis, MD 21402

Cary Presser (2)  
Process Measurements Div.  
National Institute of Standards & Tech.  
Bldg. 221, Room B312  
Gaithersburg, MD 20899

Thomas A. Pucik  
Pucik Consulting Services  
13243 Warren Avenue  
Los Angeles, CA 90066-1750

Sheldon Reaven  
Department of Technology & Society  
343-A Harriman Hall  
State University of New York  
Stony Brook, NY 11794-3760

Helen Regan  
Biology Department  
San Diego State University  
5500 Campanile Drive  
San Diego, CA 92182-4614

Patrick J. Roache  
1215 Apache Drive  
Socorro, NM 87801

F. James Rohlf  
Dept. Ecology and Evolution  
State University of New York  
Stony Brook, NY 11794

Tim Ross (2)  
Dept. of Civil Engineering  
University of New Mexico  
Albuquerque, NM 87131

Chris Roy (2)  
Dept. of Aerospace Engineering  
211 Aerospace Engineering Bldg.  
Auburn University, AL 36849-5338

Randall Ryti  
Neptune and Company  
1505 15th Street  
Los Alamos, NM 87544

J. Sacks  
Inst. of Statistics and Decision Science  
Duke University  
Box 90251  
Durham, NC 27708-0251

Len Schwer  
Schwer Engineering & Consulting  
6122 Aaron Court  
Windsor, CA 95492

Nell Sedransk (2)  
19 T. W. Alexander Drive  
P.O. Box 14006  
Research Triangle Park, NC 27709-4006

Teddy Seidenfeld  
Department of Philosophy, BH 135J  
Carnegie Mellon University  
Pittsburgh, PA 15213

Tom I-P. Shih  
Dept. of Aerospace Engineering  
2271 Howe Hall, Room 1200A  
Iowa State University  
Ames, IA 50011-2271

Don Simons  
Northrop Grumman Information Tech.  
222 W. Sixth St.  
P.O. Box 471  
San Pedro, CA 90733-0471

Munir M. Sindir  
Boeing-Rocketdyne Propulsion & Power  
MS GB-11  
P. O. Box 7922  
6633 Canoga Avenue  
Canoga Park, CA 91309-7922

Paul Slovic  
Decision Research  
1201 Oak Street  
Eugene, OR 97401

Bill Spencer (2)  
Civil Engineering & Geological Sciences  
University of Notre Dame  
Notre Dame, IN 46556-0767

Eugene Stakhiv  
U.S. Army Corps of Engineers  
Institute for Water Resources  
Casey Building  
7701 Telegraph Road  
Alexandria, VA 22315

Fred Stern  
Professor Mechanical Engineering  
Iowa Institute of Hydraulic Research  
The University of Iowa  
Iowa City, IA 52242

Scott D. Sudweeks  
Environmental Protection Agency  
Sam Nunn Atlanta Federal Center  
61 Forsyth Street, SW  
Atlanta, GA 30303-8960

Glenn Suter  
US Environmental Protection Agency  
Nat. Center for Environmental Assessmt  
26 W Martin Luther King Dr, MC-A-130  
Cincinnati, OH 45267

Ben Thacker  
B-128  
Southwest Research Institute  
6220 Culebra Road  
P. O. Drawer 28510  
San Antonio, TX 78284-0510

Kimberly Thompson  
Harvard School of Public Health  
617 Huntington Avenue  
Boston, MA 02115

Ronald R. Yager  
Machine Intelligence Institute  
Iona College  
715 Noth Avenue  
New Rochelle, NY 10801

Ren-Jye Yang  
Ford Research Laboratory  
MD2115-SRL  
P.O.Box 2053  
Dearborn, MI 4812

Simone Youngblood  
The Johns Hopkins University  
Applied Physics Laboratory  
11100 Johns Hopkins Road  
Mail Stop 13N409  
Laurel, MD 20723

## Foreign Distribution

Tom Aldenberg  
RIVM  
PO Box 1  
NL-3720 BA Bilthoven  
NETHERLANDS

Yakov Ben-Haim (2)  
Department of Mechanical Engineering  
Technion-Israel Institute of Technology  
Haifa 32000  
ISRAEL

Dr.-Ing. Michael Beer  
Assistant Professor  
National University of Singapore  
Department of Civil Engineering  
BLK E1A, #05-04, 1 Engineering Drive 2  
Singapore 117576

Mark Burgman  
School of Botany  
University of Melbourne  
Victoria, 3010  
AUSTRALIA

Roger M. Cooke  
Delft Institute of Applied Mathematics  
Delft University of Technology  
PO Box 5031, 2600 GA, Delft  
Mekelweg 4, 2628 CD, Delft  
THE NETHERLANDS

Frank Coolen  
Department of Mathematical Sciences  
Durham University  
Science Laboratories  
South Road  
Durham DH1 3LE  
UNITED KINGDOM

Gert de Cooman (2)  
Universiteit Gent  
Onderzoeksgroep, SYSTeMS  
Technologiepark - Zwijnaarde 9  
9052 Zwijnaarde  
BELGIUM

Peter Craig  
Department of Mathematical Sciences  
University of Durham  
South Road  
Durham DH1 3LE  
UNITED KINGDOM

Mark Crane  
Watts & Crane Associates  
23 London Street  
Faringdon, Oxfordshire SN7 7AG  
UNITED KINGDOM

Luis Eca (2)  
Instituto Superior Tecnico  
Department of Mechanical Engineering  
Av. Rovisco Pais  
1096 Lisboa CODEX  
PORTUGAL

Dominique Guyonnet  
BRGM-Environnemental Department  
3 avenue Claude Guillemin  
BP 6009  
45060 Orléans Cedex 2  
FRANCE

Andy Hart  
Central Science Laboratory  
Sand Hatton  
York YO41 1LZ  
UNITED KINGDOM

Charles Hirsch (2)  
NUMECA International  
Avenue Franklin Roosevelt, No. 5  
B-1050 Brussels  
BELGIUM

Antonio Irpino  
Via Masucci 2  
I-83100 Avellino AV  
ITALY

Carlo N. Lauro  
Dipartimento di Matematica e Statistica  
Università di Napoli "Federico II"  
via Cintia 26, 80126 Napoli  
ITALY

David Moens  
K. U. Leuven  
Dept. of Mechanical Engineering, Div.  
PMA  
Kasteelpark Arenberg 41  
B – 3001 Heverlee  
BELGIUM

Bernd Möller  
Institut für Statik und Dynamik der  
Tragwerke  
Technische Universität Dresden  
01062 Dresden  
GERMANY

Dwayne Moore  
Cantox Environmental  
1550A Laperriere Avenue, Suite 103  
Ottawa, Ontario  
CANADA K1Z 7T2

Ingo Neumann  
Geodetic Institute  
University Hannover  
Nienburger Strasse 1  
D-30167 Hannover  
GERMANY

Kung Ng  
Computational Dynamics Ltd.  
200 Shepherds Bush Road  
London W6 7NY  
UNITED KINGDOM

Nina Nikolova - Jeliaskova  
Institute of Parallel Processing  
Bulgarian Academy of Sciences  
25a "acad. G. Bonchev" str.  
Sofia 1113  
BULGARIA

K. Papoulia  
Inst. Eng. Seismology & Earthquake  
Engineering  
P.O. Box 53, Finikas GR-55105  
Thessaloniki  
GREECE

Dominique Pelletier  
Genie Mecanique  
Ecole Polytechnique de Montreal  
C.P. 6079, Succursale Centre-ville  
Montreal, H3C 3A7  
CANADA

Willem Roelofs  
Central Science Laboratory  
Sand Hatton  
York YO41 1LZ  
UNITED KINGDOM

Marco Zaffalon  
IDSIA  
Galleria 2, CH-6928 Manno (Lugano)  
SWITZERLAND



**Department of Energy  
Laboratories**

Los Alamos National Laboratory (11)  
Mail Station 5000  
P.O. Box 1663  
Los Alamos, NM 87545

Attn: Mark C. Anderson, MS T080  
Scott Doebling, MS T080  
S. Eisenhower, MS K557  
Francois Hemez, MS F699  
Karen Hench, MS P946  
David Higdon, MS F600  
Cliff Joslyn, MS B265  
Karen I. Pao, MS B256  
William Rider, MS D413  
Mandy Rutherford, MS T080  
Alyson G. Wilson, MS F600

Lawrence Livermore National  
Laboratory (7)  
7000 East Ave.  
P.O. Box 808  
Livermore, CA 94551

Attn: Frank Graziani, MS L-095  
Henry Hsieh, MS L-229  
Richard Klein, MS L-023  
Roger Logan, MS L-125  
Cynthia Nitta, MS L-096  
Joe Sefcik, MS L-160  
Carol Woodward, MS L-561

Idaho National Laboratory (3)  
2525 Fremont Ave.  
P.O. Box 1625  
Idaho Falls, ID 83415

Attn: R.W. Johnson, MS 5226  
Dana A. Knoll, MS 3855  
Richard R. Schultz, MS 3890

## Sandia Internal Distribution

1	MS 1179	1341	L. Lorence	1	MS 0139	1902	P. Yarrington
1	MS 1318	1411	J. R. Stewart	2	MS 0139	1904	M. Pilch
1	MS 1318	1411	M. S. Eldred	1	MS 1393	1911	K. F. Alvin
1	MS 1318	1411	D. M. Gay	1	MS 0427	2118	R. A. Paulsen
1	MS 1318	1411	L. P. Swiler	1	MS 0427	2118	S. E. Klenke
1	MS 0370	1411	T. G. Trucano	1	MS 0453	2120	M. R. Sjulín
1	MS 0378	1431	R. M. Summers	1	MS 0751	6315	L. S. Costin
1	MS 1322	1435	J. B. Aidun	1	MS 0751	6315	R. M. Brannon
1	MS 0384	1500	A.C. Ratzel	1	MS 0751	6315	A. F. Fossum
1	MS 0824	1500	T.Y. Chu	1	MS 0757	6442	G. D. Wyss
1	MS 0825	1510	M. R. Prairie	1	MS 0776	6783	R. J. MacKinnon
1	MS 0834	1512	R. D. M. Tachau	1	MS 0778	6783	P. N. Swift
1	MS 0834	1512	S. J. Beresh	1	MS 0776	6783	R. P. Rechard
1	MS 0825	1515	D. W. Kuntz	1	MS 9105	8229	A. R. Ortega
1	MS 0557	1521	D. B. Clauss	1	MS 9154	8244	R. E. Oetken
1	MS 0555	1522	R. A. May	1	MS 9042	8776	E. P. Chen
1	MS 0821	1530	A.L. Thornton	1	MS 9159	8962	H. R. Ammerlahn
1	MS 1135	1532	S. R. Tieszen	1	MS 9159	8962	P. D. Hough
1	MS 1135	1532	J. T. Nakos	1	MS 9159	8962	M. L. Martinez-
1	MS 1135	1534	J. L. Cherry				Canales
1	MS 1135	1535	M. T. Valley	1	MS 0830	12335	K. V. Diegert
1	MS 0380	1540	H. S. Morgan	1	MS 0829	12337	J. M. Sjulín
1	MS 0828	1544	A. A. Giunta	1	MS 0829	12337	S. V. Crowder
1	MS 0779	1544	J. C. Helton	1	MS 0829	12337	B. M. Rutherford
5	MS 0828	1544	W. L. Oberkampf	1	MS 0829	12337	F. W. Spencer
1	MS 0557	1544	T. L. Paez	2	MS 9018	8944	Central Technical
1	MS 0828	1544	V. J. Romero				Files
1	MS 1152	1652	M. L. Kiefer	2	MS 0899	4536	Technical Library
1	MS 0139	1900	A. Hale				